A two-level Drive - Response Model of Instationary Speech Signals

Friedhelm R. Drepper

Zentralinstitut für Elektronik, Forschungszentrum Jülich GmbH, Postfach 1913,
52425 Jülich, Germany
f.drepper@fz-juelich.de
28.01.2005

Abstract. The transmission protocol of voiced speech is hypothesized to be based on a fundamental excitation or drive process, which synchronizes the vocal tract excitation on the transmitter side and evokes the loudness and pitch perception on the receiver side. The fundamental drive can be extracted from the speech signal by using a voice-specific subband decomposition. When used as fundamental drive of a two-level drive - response model with stationary coupling on both levels, the instationary drive is able to describe instationary speech as secondary response. For simplicity each subband specific primary response is assumed to be restricted to a nonlinear synchronisation manifold. Whereas the extraction of a physiologically interpretable fundamental phase is limited to voiced sections of speech, the fundamental amplitude can as well be used for the time scale separation of unvoiced sections.

1 Introduction

The well known source and filter model describes a speech signal as linear response of an acoustic excitation, where the linear filter describes the signal forming properties of the (extended) vocal tract (Fant 1960, Schroeder 1999). In the case of voiced sections of speech the acoustic excitation results either directly from the pulsatile airflow through the vocal fold or from secondary turbulent excitation in the vicinity of constrictions of the vocal tract. The coupled dynamics of the pulsatile airflow and the glottal tissue is characterized by a dominant direction of interaction, such that the glottal oscillator can be interpreted as the driving subsystem. This encourages a description of the voiced excitation as primary response of a fundamental drive, which oscillates in the frequency range of the pitch. The resulting two-level drive-response model can be seen as a natural extension of the classical source and filter model. In contrast to the secondary response, the coupling of the primary response has to be assumed to be a nonlinear one, because the subband excitations of voiced phones are characterized by phoneme and speaker specific mode locking pheno-

mena. The description of the primary response can be simplified, by assuming the subband specific excitation dynamics of voiced phones to be restricted (entrained) to centre manifolds (Haken 1983) or synchronization manifolds (Rulkov et al. 1995).

2 Extraction of the fundamental drive

Being relevant to speech-acoustics as well as to psycho-acoustics, the fundamental drive is assumed to be characterized by the following features:

- The fundamental drive process is uniquely described by two (response related) state variables which can be chosen as fundamental amplitude and phase (Drepper 2003, 2004).
- The fundamental drive process can be extracted selfconsistently from the speech signal by using a bandpass filter decomposition of the speech signal with harmonically spaced centre filter frequencies and audiologically chosen bandwidths, the centre filter frequencies being adapted with high precision to the momentary (or recently encountered) frequency of the drive.
- In the case of voiced speech segments there exist at least two subbands, the phases of which are linearly related to the fundamental phase and which can thus be used for a reconstruction of the fundamental phase (Drepper 2003, 2004).
- The selection of the appropriate subbands can be achieved with the help of a phase synchronization criterion.
- The phase synchronization can be interpreted as a confirmation of the topological equivalence of the reconstructed fundamental drive to a glottal master oscillator in the frequency range of the pitch, which enslaves (entrains) the faster degrees of freedom of the phonation or excitation process (Kantz and Schreiber 1997, Drepper 2003, 2004).
- For voiced segments the phase velocity of the reconstructed fundamental drive can be used to improve the centre filter frequencies – a feature which can be used to construct a fast converging pitch tracker with a high time and frequency resolution.
- The amplitude of the fundamental drive can be extracted from the speech signal by using the hypothesis that the perception of loudness is related to the amplitude of the fundamental drive and
- the subjective pitch perception is closely related to the objective frequency (phase velocity) of the fundamental drive or equivalently to the frequency of the glottal master oscillator of voiced excitation.

An essential feature of the above hypothesis is the fact that the centre frequencies of the bandpass filters are spaced evenly at the lower frequency end. This is in contrast to the well known audiological filterbanks which are spaced evenly on an audiological scale, e.g. on the ERBscale (equivalent rectangular bandwidth scale) as defined by Patterson (1987). A well known implementtation of such an audiological filterbank is based on 4^{th} order complex gammatone bandpass filters with a constant analysis - synthesis delay as described

in Hohmann (2002). To be useful for the reconstruction of the fundamental drive the audiological ERB scale has been replaced by a piecewise linear and logarithmic equivalent rectangular bandwidth scale, which is exactly harmonic (linearly spaced) at the lower frequency end and logarithmically spaced at the higher frequency end (Drepper 2004). The spacing in the logarithmic range is chosen as 4 filters per octave. In this case the continuity of the ERB can be achieved for 5 (effectively 6) separable subbands. First results point into the direction that 6 subbands are sufficient to extract at least a useful first approximation of the fundamental phase. When the centre filter frequency of the first subband is denoted by F_1 and the maximal subband index is denoted by *N*, the equivalent rectangular bandwidths and other centre frequencies are given as

$$ERB_j = \begin{Bmatrix} F_1 \\ 2^{(j-5)/4} F_1 \end{Bmatrix} \quad \text{for} \quad \begin{Bmatrix} 1 \le j \le 5 \\ 5 < j \le N \end{Bmatrix},$$

$$F_j = \begin{Bmatrix} j F_1 \\ 5 \cdot 2^{(j-5)/4} F_1 \end{Bmatrix} \quad \text{for} \quad \begin{Bmatrix} 1 \le j \le 6 \\ 6 < j \le N \end{Bmatrix}.$$

The extraction of the fundamental amplitude is based on the assumption, that human auditive perception incorporates useful information on the dynamics of important sound sources of the human environment in particular on human speech. The relevant features of the loudness perception concern the scaling of the loudness as function of the signal amplitude and the relative weight of the partial loudnesses of individual subbands (Moore 1989). Using an unconventional normalisation, the dependence of the loudness L_t at time *t* on the (pressure) amplitudes $A_{i,t}$ of a set of subbands with indizes $1 \le i \le N$ can be expressed as

$$L_t = \sum_{i=1}^{N} (g_i \overline{A}_{i,t})^{\nu} \quad \text{mit} \quad \sum_{i=1}^{N} g_i^{\nu} = 1,$$

with the subband specific weigths g_i (which are proportional to inverse hearing thresholds) and average amplitudes $\overline{A}_{i,t}$. Older sources (Zwicker und Feldtkeller 1967, Moore 1989) give an exponent $\nu = 0.6$. Sottek (2000) cites newer measurements, resulting in an exponent in the range of $\nu = 0.30$. The latter value has been assumed in the study. The weights g_i can e.g. be obtained from iso-loudness contours for loudnesses above 40 phon, a range, which can be assumed to be typical for speech communication. In this loudness range the weights can roughly be assumed to scale according to a power law as function of the harmonic number h_i. Denoting the exponent as μ, we obtain $g_i \approx h_i^{\mu}$. The two mentioned older sources result in an exponent either in the range $\mu = 1$ or $\mu = 2$. The value $\mu = 1$ has the nice property that the typical amplitude spectrum of a voiced excitation (with a roughly triangular profile) is compensated (Schoeder 1999) and has thus been preferred. The drive amplitude A_t is obtained by assuming it to represent a linear homogenous function of the subband amplitudes $\overline{A}_{i,t}$.

$$A_t \quad = \quad \left(\sum_{i=1}^{N} (g_i \, \overline{A}_{i,t})^{\nu} \right)^{\frac{1}{\nu}} \quad .$$

The feasibility of the extraction of the fundamental drive and the validity of its interpretation as a reconstruction of a glottal master oscillator in the frequency range of the pitch is demonstrated with the help of a simultaneous recording of a speech signal and an electro-glottogram obtained from the (extremely useful) publicly accessible pitch analysis database of the Keele University, (ftp.cs.keele.ac.uk/pub/pitch). The upper trace of figure 1 opens an analysis window of 45 ms for a speech segment taken from the /w/ in wind spoken by the first male speaker as part of the sentence “The North wind and the sun were disputing …”. The lower trace shows the reconstruction of the fundamental phase, based on the subbands with the harmonic numbers 2, 3 and 5. (The fundamental phase is given in the wrapped form, normalized by π.) The near perfectly linear phase synchronization of these subbands, which is used for the reconstruction of the drive, is demonstrated in figure 3, which shows the corresponding phase relations to the fundamental phase. Figure 2 shows the analogues of figure 1 obtained from the simultaneous recording of the electro-glottogram. In the latter case the first four separable subbands are perfectly suited to reconstruct the fundamental phase (not shown). The comparison of figures 1 and 2 underlines the equivalence and exchangeability of the two fundamental phases. However it is important to note, that even with six perfectly linear related subbands it is not possible to determine a unique initial phase of the glottal oscillator. A second demonstration of the exchangeability of the two drives is given in figures 4 and 5, which show the result of the pitch extraction for 100 successive analysis windows covering the section “North wind and the sun”. In each figure the upper trace indicates the number of phase synchronous subbands, which are used for the reconstruction of the fundamental phase and the lower trace indicates the adapted centre filter frequency of the fundamental subband, which – in the case of voiced segments – coincides with the average phase velocity of the fundamental drive.

3 Entrainment of the primary response

In spite of the mentioned indeterminacy of the initial phase, the described reconstruction of the fundamental glottal oscillator can be used to reconstruct a speech signal. The key to this reconstruction is a two level drive – response model, which can be seen as a natural extension of the (subband version of the) well known source and filter model. The additional subsystems represent the subband specific vocal tract excitations as primary responses of the fundamental drive (Drepper, 2003-2005). The secondary drive – response subsystems describe the signal forming, which results from resonance and reverberation in the vocal

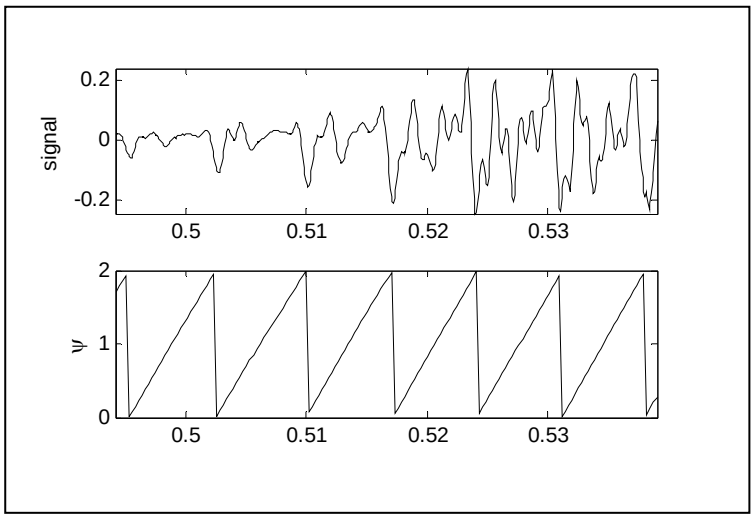

Fig. 1: The upper trace opens the analysis window for a section of a speech signal, which was taken from a /w/ like in "wind", as part of a publicly accessible speech data base, described in the text. The lower trace shows a reconstruction of the phase of the fundamental drive, obtained from the subbands 2, 3 and 5. Phases are given in units of π. The time scale corresponds to the original one and is given in units of seconds.

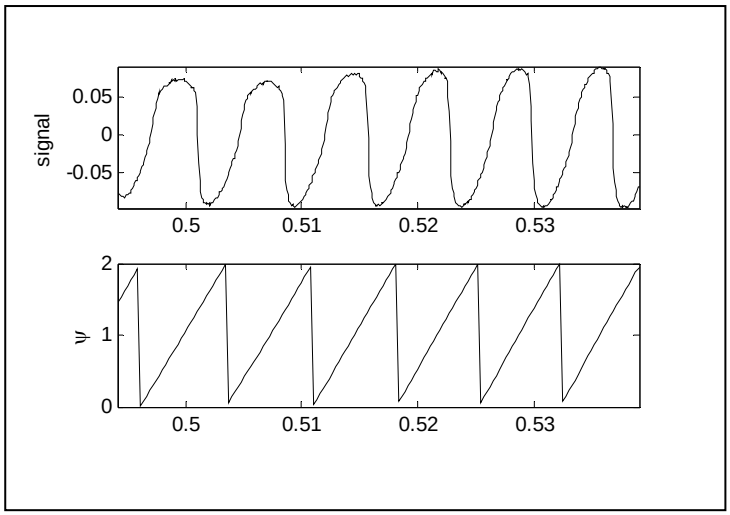

Fig. 2.: The upper trace opens the analysis window for the simultaneous section of an electro-glottogram, which was recorded simultaneously with the speech signal of figure 1. The corruption of the zero level has been reduced by applying a 75 Hz (moving average) high pass filter. The lower trace shows the reconstruction of the phase of the fundamental drive, obtained from the subbands 1 to 4.

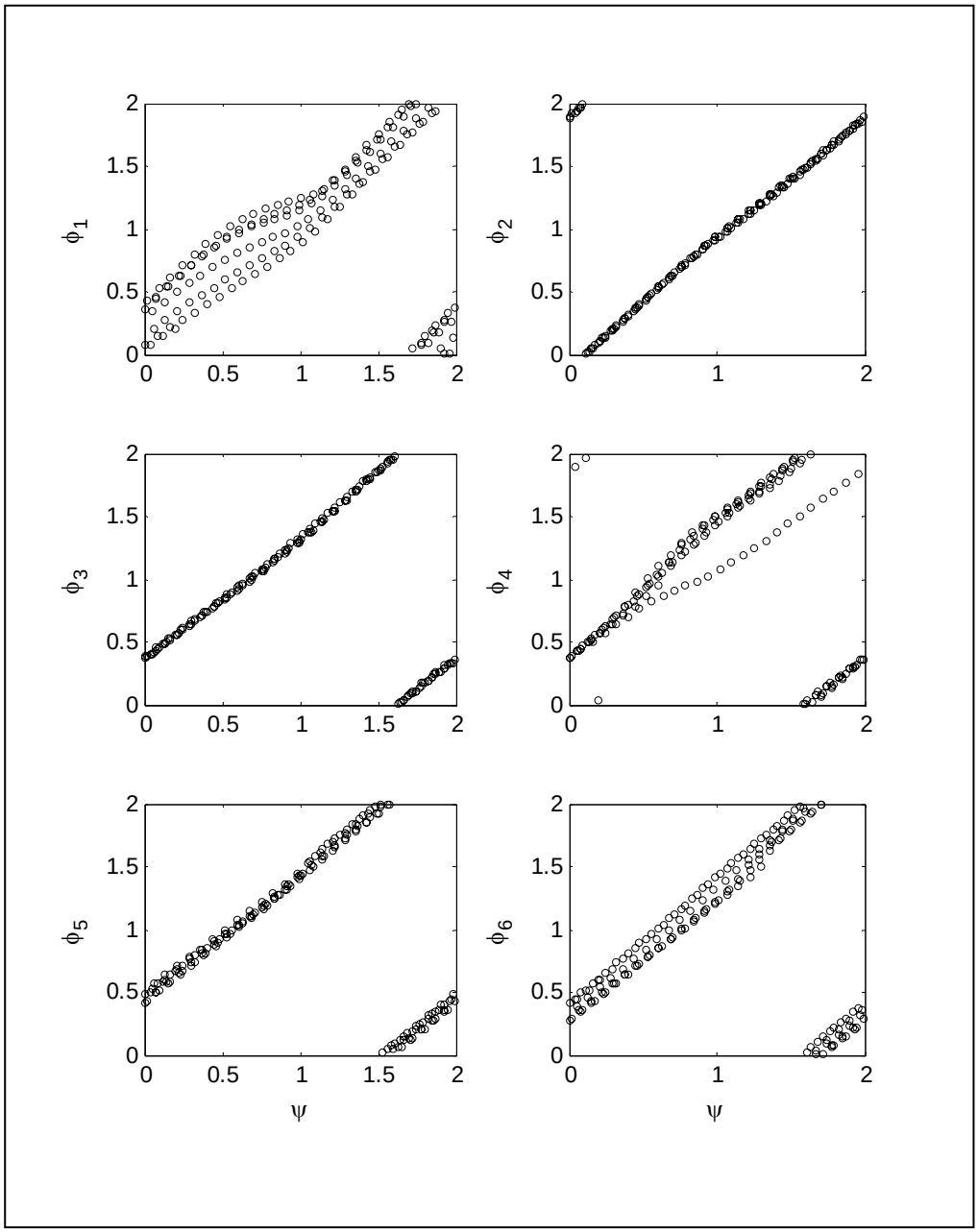

Fig. 3. The relation of the subband phases Φ_j, $(j = 1,2,...,6)$ of the speech signal of figure 1 to the fundamental phase ψ. Note that the subband phases Φ_j are given in a partially unwrapped form, depending on the harmonic number of the respective subband. The enlarged range of the subband phases is finally normalized by the respective harmonic number. The subbands 2, 3 and 5 are characterized by near perfectly linear phase relations, whereas the other subbands express the obvious aperiodicity of the signal in figure 1.

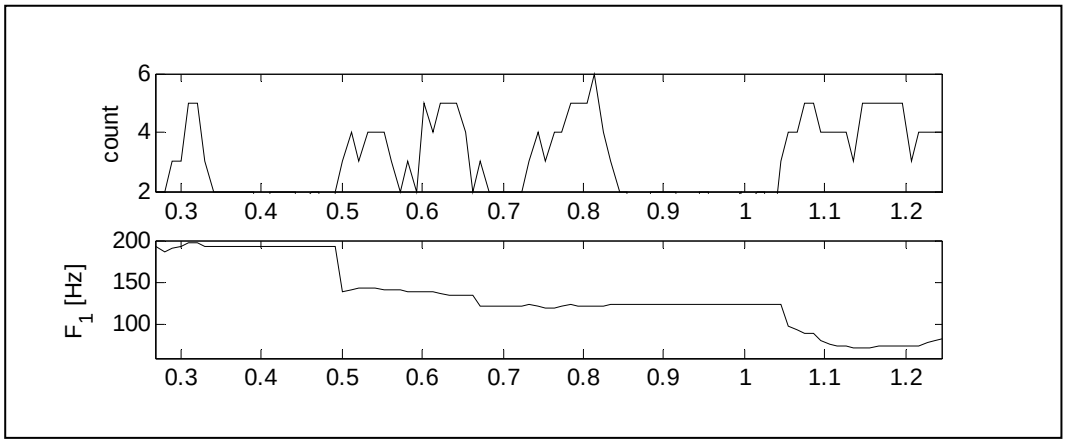

Fig. 4. Results of the pitch tracking for 100 analysis windows as described in figure 1. The upper trace indicates the count of the number of near perfectly linear phase relations, which have been uncovered by the precise adjustment of the centre filter frequencies. The lower trace indicates the adjusted centre filter frequency of the fundamental subband. For unvoiced windows the filter frequency of the most recent voiced window is indicated. The time scale corresponds to the one of figures 1 and 2.

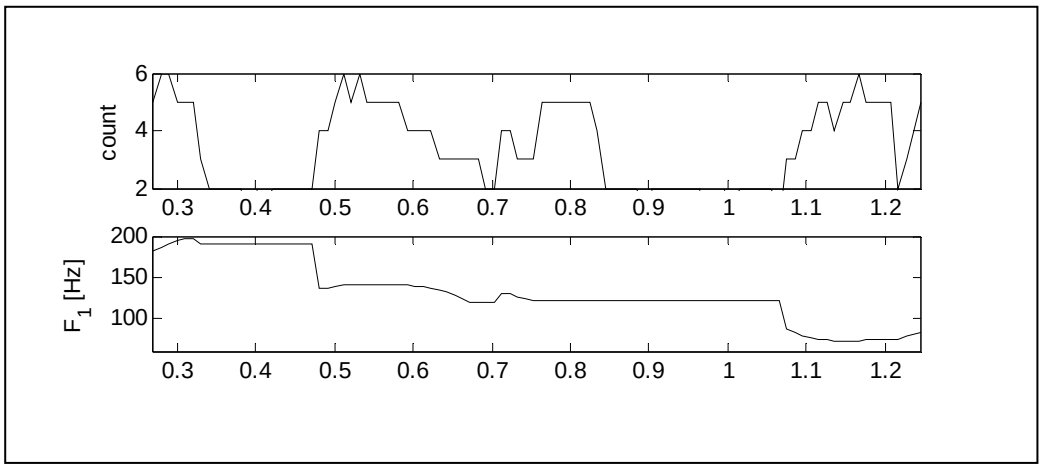

Fig. 5. Results of the pitch tracking for 100 analysis windows as described in figure 2. The comparison with figure 4 shows that the count of the number of perfectly phase synchronous subbands is generally higher in figure 5 in spite of the fact, that the electro-glottogram is corrupted by fluctuations of the zero level. The good agreement of the two centre filter frequencies supports the interpretation of the fundamental drive in terms of a fundamental glottal master oscillator of voiced excitation.

tract. Both effects are known to be represented well by linear response models (Vary et al. 1998, Schroeder 1999). The introduction of a nonlinear coupling of the subband excitations to a common fundamental drive represents an easy option to describe the cross-correlation between the subbands.

In addition to this the introduction of the fundamental drive opens the possibility of a more subtle time scale separation (Drepper 2005), which is crucial for the time series analysis of instationary speech signals. The usual time scale separation is based on the assumption of a (weakly or wide sense) stationary excitation (and constant vocal tract filter) throughout the extent of the analysis window, usually chosen not larger than 20 ms, which is in contrast to the average length of a phoneme of about 100 ms. The introduction of the fundamental drive as an additional level of the time series model has the advantage that the assumption of stationarity can now be restricted to the drive - response couplings and that the explicitly reconstructed fundamental drive is freed from any stationnarity assumption. The dropping of the latter assumption opens the possibility to extend the analysis window up to more than 40 ms (depending e.g. on the gender or average pitch of the speaker).

For simplicity the nonlinear drive – response dynamics of each subband specific excitation subsystem is assumed to be restricted (entrained) to a centre-manifold in the combined state space of drive and response (Haken 1983), which (in the simple case) can be interpreted as a synchronization manifold (Rulkov et. al. 1995). The synchronisation manifold represents a two-dimensional surface in the four-dimensional state space, which relates the state of the primary response uniquely to a two dimensional state of the fundamental drive (Drepper 2003-2005). This assumption appears to be well justified at least for the lower harmonic subbands, since for these subbands the coupling of the excitation to the fundamental glottal oscillator is comparatively fast compared to the time scale of the secondary response dynamics. In any case, the assumption of so called generalized synchronization in a drive – response system (Rulkov et al. 1995) greatly simplifies the parameter estimation in comparison to the one of a fully dynamic nonlinear response model.

As part of the time scale separation each subband specific synchronization manifold is assumed to take the form of the product of a slowly varying univariate function of the fundamental amplitude and a potentially fast varying function of the fundamental phase. In accordance to the common linear source and filter model, the function of the amplitude is assumed to be linear. The remaining univariate phenomenological function of the fundamental phase is denoted as coupling function. To facilitate the determination of phases, all oscillator states are described by complex variables. Assuming formant resonances with well separated time scales (transfer functions with isolated poles) and temporarily neglecting the (nasal) reverberation in the vocal tract, the two level drive - response system of subband j is described by the following complex conditional stochastic process (nonlinear drive- response model)

$$X_{j,t+\Delta} \;=\; b_j X_{j,t} \;+\; A_t G_{j,p}(\psi_t) \;+\; A_t \sigma_j \xi_{j,t}$$

with $$G_{j,p}(\psi_t) \;=\; \sum_{k \in S_{j,p}} c_{j,k} \exp(i k \frac{\psi_t}{p}) \;,$$

where $X_{j,t}$ denotes the complex output of the filter bank, Δ the subband specific predict-tion step length, b_j the subband specific complex resonator parameter, ψ_t the fundamental phase at time *t*, $\xi_{j,t}$ a (0,1) Gaussian white noise process and σ_j the (time independent) subband specific part of the standard deviation. The coupling function $G_{j,p}(\psi_t)$ comes in several qualitatively different variants. The first variant represents a 2π periodic function. This case corresponds to the described unique function of the state of the drive and is thus suited to describe (simple) vowel or nasal type signals or subbands. The other two variants depend necessarily on the unwrapped fundamental phase. They either have a period length $2\pi\, p$ (with integer *p*) which approximates the analysis window length or an intermediate period length. In the case of the full period length, the dependence on the fundamental phase can be interpreted as a near perfect substitute for a dependence on time. Thus the larger period coupling function is suited to describe a general excitation of the usual source and filter model. This coupling function is thus suited to analyse voiced consonants and the continuous transition from voiced phones to unvoiced ones. The coupling functions with intermediate period lengths are expected to be useful for the analysis of the micro tremor, which is supposed to be of special importance for speaker recognition. Thus at a closer look the (1:n) synchronization of vowel type subbands may have to be replaced by an (m:n) entrainment. For all variants of the coupling function the estimation of the Fourier coefficients $c_{j,k}$ (and resonator parameter b_j) can be reduced to multiple linear regression (Drepper 2003-2004).

4 Outlook

Vowels and nasals are characterized by the fact, that the time points of the glottal closure events can be extracted from the speech signal (Vary et al. 1998, Schroeder 1999). For these phonemes the drive – response reconstruction of the excitation can be used to upgrade the equivalence of the fundamental drive to the glottal master oscillator, by using these time points, to determine the initial phase of the fundamental drive. Thus the described two level source and filter model opens the possibility, to relate the well known classes of voiced phonemes as well as different speakers to qualitatively different features of the reconstruct-ted drive – response dynamics. This creates new hope that the supervised learning of phoneme and speaker recognition can substantially be shortened. The voice specific time scale separation opens a new perspective to solve the notoriously hard cocktail party problem. The innovative concept (based on “reengineering” of the entrance level of the human auditive pathway) to use self-consistently determined fundamental drive specific filterbanks to replace single level time series models by two level ones, may also turn out to

be useful in other areas of science, which are so far obscured by a complex mixture of instationarity and nonlinearity.

The author would like to thank V. Hohmann, B. Kollmeier and J. Nix, Oldenburg, M. Kob, C. Neuschaefer-Rube, C.Hoelper and P. Vary, Aachen, G. Langner, Darmstadt, N. Stollenwerk, London, P. Grassberger, H. Halling, M. Schiek and P. Tass, Jülich for helpful discussions.

References

Drepper F.R., *Fortschritte der Akustik-DAGA'03*, (2003)
Drepper F.R. in C. Manfredi (editor), *MAVEBA 2003*, Firenze University Press (2004)
Fant G. *Acoustic theory of speech production*, Mouton, 'S-Gravenhage (1960)
Haken H., *Advanced Synergetics*, Springer, Berlin (1983)
Hohmann V., *Acta Acustica* **10**, 433-442 (2002)
Kantz H., T. Schreiber, *Nonlinear time series analysis*, Cambridge University Press (1997)
Moore B.C.J., *An introduction to the Psychology of hearing*, Academic Press (1989)
Patterson R.D., *J.Acoust.Soc.Am*. **82**, 1560-1586 (1987)
Rulkov N.F. , M.M. Sushchik, L.S. Tsimring, H.D.I. Abarbanel, *Phys. Rev. E* **51**, 980-994 (1995)
Schroeder M.R., *Computer Speech*, Springer (1999)
Sottek R., *Modelle zur Signalverarbeitung im menschlichen Gehör*, Verlag M. Wehle, Witterschlick/Bonn (1993)
Vary P., U. Heute, W. Hess, *Digitale Sprachsignalverarbeitung*, B.G. Teubner Verlag, Stuttgart (1998)
Zwicker E. und Feldtkeller R., *Das Ohr als Nachrichtenempfänger*, Hirzel Verlag, (1967)