A Two-Level Drive – Response Model of Non-stationary Speech Signals

Friedhelm R. Drepper

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

Abstract. The transmission protocol of voiced speech is hypothesized to be based on a fundamental drive process, which synchronizes the vocal tract excitation on the transmitter side and evokes the pitch perception on the receiver side. A band limited fundamental drive is extracted from a voice specific subband decomposition of the speech signal. When the near periodic drive is used as fundamental drive of a two-level drive-response model, a more or less aperiodic voiced excitation can be reconstructed as a more or less aperiodic trajectory on a low dimensional continuous synchronization manifold (surface) described by speaker and phoneme specific coupling functions. In the case of vowels and nasals the excitation can be described by a univariate coupling function, which depends on the momentary phase of the fundamental drive. In the case of other voiced consonants the coupling function may as well depend on a delayed fundamental phase with a phoneme specific time delay. The delay may exceed the length of the analysis window. The resulting long range correlation cannot be analysed or synthesized by models assuming stationary excitation.

1 Introduction

Speech signals are known to contain obviously non-stationary segments, which constitute a cue for stop consonants and which are characterized by isolated, non repetitive events with a duration of less than a couple of pitch periods. The present study is focussed on segments of speech, which cannot easily be classified as non-stationary, in particular on sustained voiced segments, which are characterized by repetitive time pattern. The vocal tract excitation of voiced speech is generated by a pulsatile airflow, which is strongly coupled to the oscillatory dynamics of the vocal fold. The excitation is created immediately in the vicinity of the vocal fold and/or delayed in the vicinity of a phoneme specific constriction of the vocal tract [1-3]. As has been pointed out by Titze [4], a mechanistic model of a dynamical system suitable to describe the self-sustained oscillations of the glottis cannot be restricted to state variables of the vocal fold alone, but has to be extended by state variables of the sub- and supraglottal aerodynamic subsystems.

Due to the strong nonlinearities of the coupled dynamics non-pathological, standard register phonation dynamics is characterized by a stable synchronization or

M. Faundez-Zanuy et al. (Eds.): NOLISP 2005, LNAI 3817, pp. 125 – 138, 2005.

mode locking of several oscillatory subsystems including the two vocal folds. The synchronization can furthermore be assumed to have the effect that some of these subsystems become topologically equivalent oscillators, whose states are one to one related by a continuous mapping with a continuous invers (conjugation) [5, 6]. Due to the pronounced mass density difference of about 1:1000 the coupling between the airflow and the glottal tissue is characterized by a dominant direction of interaction, such that the glottal oscillators can affirmatively be assumed to be a subset of those topologically equivalent oscillators. The (conjugation type) synchronization of the vocal folds has been described by kinematic and dynamic models [7, 8]. The glottal oscillators can be used to define a single glottal master oscillator, which enslaves (synchronizes) or drives the other oscillatory degrees of freedom including the higher frequency acoustic modes.

Time series of the electro-glottogram or of the sound pressure signal can more or less safely be used to reveal an oscillator, which is topologically equivalent to the glottal master oscillator. In the case of nonpathological voiced speech both types of observation reveal a unique frequency of voiced phonation, the so called fundamental frequency, which is also known to have a perceptional counterpart, the pitch. As has already been observed by Seebeck [9], human pitch perception does not rely on spectral components of the speech signal in the frequency range of the fundamental frequency. In spite of numerous attempts, the extraction of the momentary fundamental frequency out of the speech signal has not yet reached the generality, precision and robustness of auditive perception and of the analysis based on the electro-glottogram [2, 10 -12].

The time series of successive cycle lengths of oscillators, which are (implicitly assumed to be) equivalent to the glottal master oscillator show an aperiodicity with a wide range of relevant frequencies reaching from half of the pitch down to less than 0.1 Hz. Except at the high frequeny end the deviation of the glottal cycle lengths from the long term mean forms a non-stationary stochastic process. More or less distinct frequency bands or time scales have been described as: subharmonic bifurcation [8], jitter, microtremor and prosodic variation of the pitch [7, 12]. As a general feature, cycle length differences increase with the time scale, the relative differences ranging from less than 1 % up to more than 30%. In spite of the partially minor amplitudes of aperiodicity all or most of these frequency bands appear to be perceptionally relevant [13]. Some of them are known to play a major role for the non-symbolic information content of speech.

The relevant frequency range of the excitation of voiced speech extends at least one order of magnitude higher than the fundamental frequency. It is therefore common practice to introduce a time scale separation, which separates the high frequency acoustic phenomena of speech signals above the pitch from the subharmonic, subacoustic and prosodic ones below the pitch. A simple approach towards time scale separation starts with the assumption of a causal frequency gap, which separates the frequency range of the autonomous lower frequency degrees of freedom from the dependent degrees of freedom (modes) in the acoustic frequency range. In the main stream approach of speech analysis this has lead to the more or less explicit assumption that the voiced (and unvoiced) excitation is wide sense stationary in the analysis window, which is usually chosen as 20 ms [2, 3]. The assumption of wide sense stationarity is closely related to the assumption that the excitation process

can be described as a sum of a periodic process and filtered white noise with a time invariant, finite impulse response filter. In the case of voiced excitation there exists multiple evidence that this assumption is not fulfilled [14, 15]. In a first step of improvement the voiced excitation has been described as stochastic process in the basin of attraction of a low dimensional nonlinear dynamical system [14, 15]. The assumption of a low dimensional dynamical system, however, is in contradiction to the observed non-stationary aperiodicity of the glottal cycle lengths.

The present study introduces an analysis of (sustained) speech signals, which does not assume a periodic fundamental drive nor an aperiodic drive, which obeys a low dimensional dynamics. The assumption of a causal frequency gap is avoided by treating the more or less aperiodic voiced broadband excitation as an approximately deterministic response of a near periodic, non-stationary fundamental drive, which is extracted continuously from voiced sections of speech with uninterrupted phonation [16, 17]. The extraction of the fundamental drive includes a confirmation that the drive can be interpreted as a topologically equivalent reconstruction of the glottal master oscillator which synchronizes the vocal tract excitation [16].

As an important property of non-pathological, standard register voiced speech the state of the fundamental drive is assumed to be described uniquely by a fundamental phase, which is related to pitch perception, and a fundamental amplitude which is related to loudness perception. Whereas the extraction of the 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. The (response related) state of the fundamental drive should not be confused with the state of the dynamical system, which describes the self-sustained oscillations of the glottis [4]. The phase of the glottal master oscillator should rather be compared with a phase, which is suited to describe a unique state on the limit cycle, which attracts the self sustained oscillations of the glottis.

As result of a detailed study of the production of vowels (with a sufficiently open vocal tract to permit the manipulation of airflow velocity sensors) Teager and Teager [18] pointed out that the conversion of the potential energy of the compressed air in the subglottal airduct to convective, acoustic and thermal energy happens in a highly organized cascade. They observed that the astonishingly complex convective airflow pattern within the vocal tract (flow separations, vortex rings, swirly vortices along the cavity walls, …) show a degree of periodicity in time, which is comparable to the one of the corresponding far field acoustic response.

Also in the case of sustained voiced fricatives (and of vowel – voiceless fricative transitions) the far field acoustic response indicates a causal connection to the glottal dynamics [19]. It is therefore plausible to assume that at least a part of the frequency range of the convective flow pattern on the upstream side of the fricative specific constriction shows a vowel type periodicity. However, there is still a lot of speculation about the relevant delays of the cause and effect relationship between the primary response and the glottal dynamics. In the case of the fricative specific retarded excitation the delay may assume a large value, due to (comparatively slow) subsonic convective transport of the relevant action (trigger). The speculation refers in particular to the question, whether the subsonic transport is limited to the downstream side of the phoneme specific constriction [19] or applies to the whole distance starting from the glottis.

It cannot be excluded that the delay (or memory) of the subsonic excitation may reach the length of the conventional analysis window of 20 ms. In this case the resulting long range correlation cannot be analysed affirmatively by conventional methods assuming stationary excitation within the analysis window. The continuous reconstruction of the glottal master oscillator for segments of uninterrupted phonation opens the possibility to describe the excitation as superposition of a direct and a delayed phase locked response with correct long range correlation.

As has also been pointed out by Teager and Teager [18] there are many reasons to assume that the human auditive pathway uses analysis tools, which deviate from spectral analysis. Teager proposed a phenomenological approach, which is based on short term analysis of the distribution of energy in different frequency bands [21]. The present approach is focussed on a phenolmenological speech production model, which extends the validity range of the classical source and filter model, which is also grounded on evidence from speech physiology and psychoacoustics and which is suited to bring additional light to the complex airflow pattern of voiced consonants, which are extremely difficult to analyse in vivo [18], in vitro [19] and in silico [19].

2 Extraction of the Fundamental Drive

The amplitude and phase of the fundamental drive are extracted from subband decompositions of the speech signal. The decompositions use complex (4th order gammatone) bandpass filters with roughly approximate audiological bandwidths ΔF and with a subband independent analysis – synthesis delay as described in Hohmann [22].

The extraction of the fundamental phase ψ_t is based on an adaptation of the best (central) filter frequencies F_j of the subband decomposition to the momentary frequency of the glottal master oscillator (and its higher harmonics) [16, 17]. At the lower frequency end of the subband decomposition the best filter frequencies F_j are centred on the different harmonics of the analysis window specific estimate of the fundamental frequency. In the next higher frequency range the best filter frequencies are centred on pairs of neighbouring harmonics.

$$F_j = \begin{Bmatrix} j F_1 \\ \sqrt{j(j-1)}\ F_1 \end{Bmatrix} \quad \text{for} \quad \begin{Bmatrix} 1 \le j \le 6 \\ 6 < j \le 12 \end{Bmatrix} \tag{1a}$$

$$\Delta F_j = \begin{Bmatrix} F_1 \\ 2\ F_1 \end{Bmatrix} \quad \text{for} \quad \begin{Bmatrix} 1 \le j \le 6 \\ 6 < j \le 12 \end{Bmatrix}. \tag{1b}$$

As a second feature of human speech it is assumed that voiced segments of speech are produced with at least two subbands, which are not distorted by vocal tract resonances or additional constrictions of the airflow [17]. In the case of subbands with separated harmonics, $1 \le j \le 6$, the absence of a distortion is detected by nearly linear relations between the unwrapped phases of the respective subband states. For sufficiently adapted centre filter frequencies such subbands show an (n:m) phase locking. The corresponding phase relations can be interpreted to result from (n:1) and (m:1) phase relations to the fundamental drive. The latter ones are used to reconstruct

the phase velocity of the fundamental drive. In the case of a subband with paired harmonics, $6 < j \le 12$, the phase relation to the fundamental drive is obtained by determining the Hilbert phase of the modulation amplitude of the respective subband.

The phase velocity of the fundamental drive is used to improve the centre filter frequencies. For voiced phones the iterative improvement leads to a fast converging fundamental phase velocity $\dot{\psi}_t$ with a high time and frequency resolution. Based on a, so far, arbitrary initial phase, successive estimates of $\dot{\psi}_t$ lead to a reconstruction of the fundamental phase ψ_t, which is uniquely defined for uninterrupted segments of confirmed topological equivalence [17]. The uninterrupted continuation of the fundamental phase can even be achieved in cases of a confirmation gap as long as there remains an overlap of confirmed analysis windows. The latter feature can e.g. be used for the analysis of vowel-nasal transitions (figures 5 and 6).

The extraction of the **fundamental amplitude** A_t 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 loudness perception concern the scaling of the loudness as function of the signal amplitude and the relative weights of the partial loudnesses of individual subbands [10]. The fundamental amplitude A_t is assumed to be related to loudness perception by a power law [17]. The exponent $1/\nu$ is chosen such that the fundamental amplitude represents a linear homogenous function of the time averaged amplitudes $\overline{A}_{j,t}$ of a synthesis suited set of subbands with approximately audiological bandwidths,

$$A_t = \left(\sum_{j=1}^{N} (g_j \overline{A}_{j,t})^{\nu} \right)^{\frac{1}{\nu}} \qquad \text{with} \qquad \sum_{j=1}^{N} g_j^{\nu} = 1 . \tag{2}$$

Zwicker, Feldtkeller [23] and Moore [10] give an exponent $\nu = 0.6$. Sottek [24] cites newer measurements, resulting in an exponent in the range of $\nu = 0.3$. The latter value has been adopted in the study. The weights g_j are proportional to inverse hearing thresholds. In the range up to 3 kHz they can be roughly approximated by the power law $g_j \approx h_j^{\mu}$, where h_j represents the (integer) centre harmonic number, which approximates the ratio F_j / F_1. The present study uses $\mu = 1$ [3, 23]. The synthesis suited set of subbands is generated by replacing the over-complete subband set $6 < j \le 12$ by a set $6 < j \le N$, which is spaced equidistantly on the logarithmic frequency scale with 4 filters per octave,

$$F_j = 5 \cdot 2^{(j-5)/4} F_1, \qquad \Delta F_j = 2^{(j-5)/4} F_1 . \tag{3}$$

The feasibility of the extraction of the fundamental drive as well as the validity of its interprettation as a reconstruction of a glottal master oscillator of voiced excitation is demonstrated with the help of simultaneous recordings of a speech signal and an electro-glottogram, which have been obtained from the pitch analysis database of Keele University [25]. The upper panel of figure 1 shows the analysis window for a segment of the speech signal, which was taken from the /w/ in the first occurrence of the word “wind” spoken by the first male speaker. The lower panel shows the reconstruction of the fundamental phase (given in wrapped up form), based on the set of separable subbands with the harmonic numbers 2, 3 and 5.

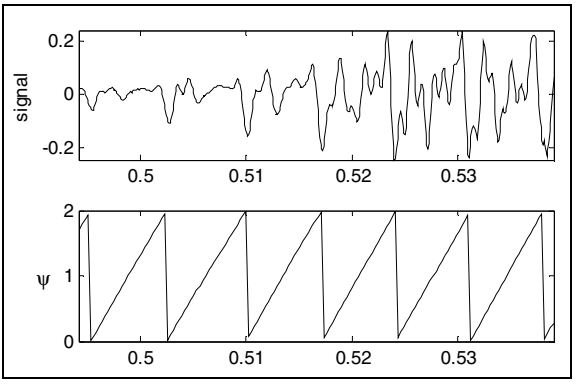

Fig. 1. Upper panel: 45 ms of a speech signal, which was taken from the /w/ in the word "wind" representing part of a publicly accessible pitch analysis data base [25]. The lower panel shows the reconstruction of the fundamental phase ψ in units of π. The time scale (in units of seconds) corresponds to the original one.

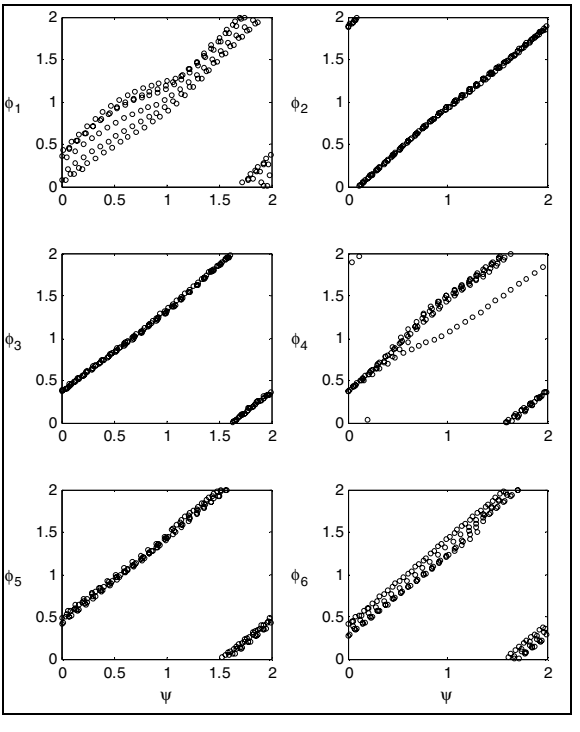

Fig. 2. Relation of the subband phases Φ_j, $(j = 1, 2, ..., 6)$, obtained from the speech signal of figure 1, to the fundamental phase ψ. The subbands 2, 3 and 5 are characterized by near perfectly linear phase relations, whereas the other subbands are found to be unsuited for the reconstruction of the fundamental phase.

The near perfectly linear phase locking of these subbands, which is used for the reconstruction of the drive, is demonstrated in figure 2. The subband phases Φ_j are given in a partially unwrapped form, depending on the respective centre harmonic number h_j . The enlarged range of the subband phases is normalized by the same centre harmonic number. Alternatively the fundamental phase can also be obtained from a subband decomposition of the electro-glottogram. The exchangeability of the two phase velocities is demonstrated in figure 3, which shows the relation between the two fundamental phases for the speech segment, which covers the "win" part of the word "wind", uttered by the first female speaker. The phase shift between the two phases did not change significantly during the 160 ms being covered.

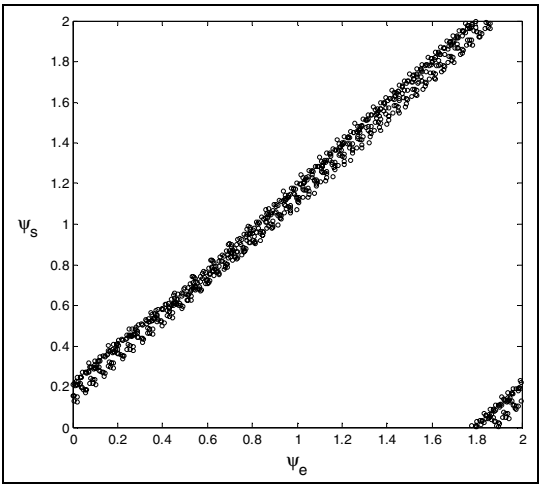

Fig. 3. Relation between the wrapped up fundamental phase ψ_s, obtained from the speech signal, and the fundamental phase ψ_e, obtained from the electro-glottogram. Both fundamental phases are extracted from 160 ms of uninterrupted voiced speech.

3 Entrainment of the Primary Response

In spite of the (temporary) arbitrariness of the initial fundamental phase, the reconstructed glottal master oscillator can be used as fundamental drive of a two level drive – response model, which is suited to describe voiced speech as secondary response [16, 26, 27]. The additional subsystem describes the excitation of the vocal tract as primary response of the fundamental drive and the classical secondary response subsystem describes the more or less resonant "signal forming" on the way through the vocal tract as action of a linear autoregressive filter, which (in a first approximation) is assumed as independent of the fundamental phase.

As a particular advantage of the two-level drive- response model the fundamental phase cannot only be interpreted as state variable of the fundamental drive. The

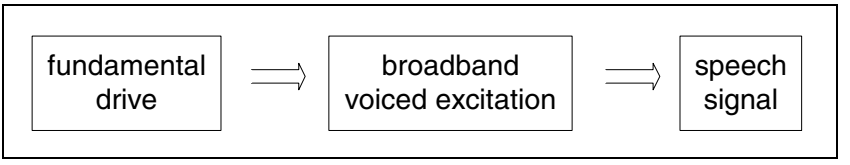

Fig. 4. The two-level drive – response model

unwrapped fundamental phase can also be assumed to be approximately proportional to time. As a characteristic simplifying assumption of the two-level drive-response model the near periodic time profile of the excitation is replaced by a precisely periodic fundamental phase profile [16, 26, 27].

In the context of the drive- response model this means that the excitation E_t at time *t* is assumed to be restricted (phase locked or entrained) to a generalized synchronization manifold (surface) in the combined state space of the fundamental drive and the primary response [28-30]. In the simplest case, the time dependence is replaced by a unique dependence on the simultaneous state of the fundamental drive [28,29]. More generally, the dependence of the primary response on the fundamental drive takes the form of a multi-valued mapping [30], which, however, can be expressed as a unique function of the unwrapped fundamental phase ψ_t,

$$E_t \quad = \quad A_t \, G_p(\psi_t) \quad = \quad A_t \sum_{k=0}^{K} c_k \, \exp(i\,k\,\frac{\psi_t}{p}) \,. \tag{4}$$

As part of the improved time scale separation the generalized synchronization manifold is assumed to be described by the product of the slowly variable fundamental amplitude A_t and the potentially fast varying complex coupling function $G_p(\psi_t)$, the real part of which describes the broadband excitation. To generate the mentioned multi-valued mapping (with unique branches) the coupling function $G_p(\psi_t)$ has to be assumed as a $2\pi\,p$ periodic function of the unwrapped fundamental phase ψ_t. The integer period number $p \geq 1$ defines the number of different branches of the multi-valued mapping. Coupling function $G_p(\psi_t)$ can thus be well approximated by the finite Fourier series of equation (4), which, however, has to be estimated for non-equidistant phases!

Voiced excitations can be represented by coupling functions with values of *p*, which are distinctly smaller than the number of fundamental cycles within the analysis window. The case $p = 1$ corresponds to the normal voice type characterized by a unique mapping and the case $p = 2$ corresponds to the period doubling voice type. The latter voice type can e.g. be identified by observing alternating period lengths, a feature, which is described as diplophonia in vocology. When *p* exceeds the number of fundamental cycles within the analysis window, equation (4) is able to describe a fully general excitation, including the unvoiced case. The real part of the coupling function of the normal voice type, $G_1(\psi_t)$, can be expressed as a polynomial in the harmonic functions $\cos(\psi_t)$ and $\sin(\psi_t)$. Similar polynomial coupling (or waveshaper) functions have been introduced by Schoentgen [31] to synthesize vowels with realistic vocal aperiodicities.

The excitation parameters c_k cannot be determined independently from the parameters which characterize the vocal tract resonances. In the standard approach the parameter estimation is performed hierarchically, by making the higher level assumption that the excitation has a nearly white (or tilted) spectrum. When dropping this assumption, special care has to be taken to avoid numerical instabilities in the case of a near periodic fundamental drive. To achieve a comparable numerical robustness, it is useful to perform separate parameter estimations for different frequency bands and to use optimally chosen (subband specific) time step lengths Δ for the autoregressive models. The bandwidths of the subbands should be chosen substantially broader than the bandwidths of the vocal tract resonances, which are most relevant for the respective subband. It is therefore advantageous to use a subband decomposition with larger bandwidths, than the ones used for the extraction of the fundamental drive. The band limitation can be used to reduce the number of resonances (poles of the autoregressive filter), which are relevant for the respective subband. Useful choices are two poles and one subband per octave or one pole and two subbands. For simplicity one pole and maximally two subbands of decomposition (1) and (3) are chosen. Interpreting excitation E_t of equation (4) as the aggregate of the set of subband specific excitations $E_{j,t}$ with subband specific coupling functions $G_{j,p}(\psi_t)$ and index sets $S_{j,p}$ of the Fourier type decomposition, we arrive at the following subband specific conditional stochastic process with a two-level drive – response model as deterministic part (skeleton) [16, 26, 27],

$$X_{j,t+\Delta} \;=\; -b_j\, X_{j,t} \;+\; A_t\, G_{j,p}(\psi_t) \;+\; A_t\, \sigma_j\, \xi_{j,t}\,, \tag{5}$$

where $X_{j,t}$ denotes the complex state of the subband with index j, b_j the complex, subband specific resonator parameter, $\xi_{j,t}$ a (0,1) Gaussian complex white noise process and $A_t\,\sigma_j$ the standard deviation, which for simplicity has been assumed to be not dependent on the fundamental phase. As an important computational advantage the estimation of the complex excitation and resonator parameters $c_{j,k}$ and b_j can be reduced to multiple linear regression. The subband specific summation index set $S_{j,p}$ in equation (4) is chosen in accordance to the respective bandpass filter. To avoid a bad conditioning of the parameter estimation in the case of near periodic driving, the index set $S_{j,p}$ is pruned by the index, which equals the respective centre harmonic number h_j. Together with the option, to extend the analysis window due to the explicit reconstruction of the non-stationary part, these precautions lead to a precise and robust reconstruction of the voiced excitation.

When the speech signal of the respective analysis window can be described successfully by model (5) with a low periodicity $p \le 2$, the speech signal has a high probability to belong to a vowel or a nasal. As is well known (and shown in figures 5 and 6) vowels and nasals are characterized by the fact that the time points of glottal closure can be detected as a unique pulse (or as a unique outstanding slope). Since there is no syllable without a vowel kernel, such kernels can be used to resolve the arbitrariness of the initial fundamental phase and to calibrate the wrapped up fundamental phase in terms of the time interval since the last glottal closure.

When the respective speech signal cannot be described successfully by a single low period coupling function, the unique reconstruction of the fundamental phase for uninterrupted segments of voiced phonation opens the possibility to extend model (5) by a retarded (subsonic) excitation which is suited to describe the delayed characteristic response of fricatives. According to the more detailed (aeroacoustic) view of speech production [18-20] the excitation of voiced fricatives (and of vowel – voiceless fricative transitions) should be extended by an additional or alternative coupling function, which depends on a delayed fundamental phase with a phoneme (and potentially speaker) specific delay τ,

$$X_{j,t+\Delta} = -b_j X_{j,t} + A_t G_{j,I}(\psi_t) + A_{t-\tau} G_{j,II}(\psi_{t-\tau}) + A_t \sigma_j \xi_{j,t} . \quad (6)$$

The average delay τ between the sonic and the subsonic excitation accounts for the additional time, which is needed for the (comparatively slow and quiet) subsonic transport of kinetic energy by convective airflow to the phoneme specific site of the vocal tract, where the enhanced transformation to acoustic (and thermal) energy takes place (typically at the teeth). Assuming a near optimal evolutionary adaptation of human speech production leading to a near optimal support of the distinction between the sonic coupling function $G_{j,I}(\psi_t)$ and the subsonic one $G_{j,II}(\psi_{t-\tau})$, a typical physiological tremor frequency of 7 Hz would correspond to a typical delay time of about 35 ms. For delay times in excess of 20 ms, the respective autocorrelation cannot be analysed by conventional methods, which assume uncorrelated excitation in non-overlapping analysis windows (of typically 20 ms length).

4 Long Range Correlation in a Vowel – Nasal Diphone

Contrary to the mainstream view, properties of the excitation can be used advantageously as additional cues for phoneme recognition. As a first example, where the long range correlation of a voiced speech signal represents a potential cue, the vowel- nasal diphone of figure 3 is selected, which represents the transition from the vowel to the nasal in the word “wind”. Due to the difference in length and shape of the nasal tract compared to the vocal tract, a transition between a nasal and a vowel can be discerned by a sudden change of the phase position of the glottal pulse [32] relative to the normal position for vowel kernels. Fortunately the shift of the glottal pulse happens so fast, that the gap of the confirmation of the topological equivalence of the fundamental drive to the glottal master oscillator is short enough to permit an uninterrupted continuation of the fundamental phase. Figures 3, 5 and 6 are obtained with a 30 ms time window of analysis and an advancement step size of 5ms. That means that the gap of the confirmation of the topological equivalence was shorter than 5 advancement steps. Figure 3 confirms the successful continuation of the fundamental phase.

Figures 5 and 6 reveal a phase shift of about 1/12 of the fundamental cycle. Knowing the fundamental frequency of 230 Hz the phase shift can be translated to a time shift of about 0.36 ms and a distance shift of about 12 cm. As has been pointed out by Kawahara and Zolfaghari [32] the time or distance shift has to be interpreted as an effective shift, which includes a group delay difference, which results from

differing vocal tract and nasal tract resonances. The latter ones are known to be increased by various sinuses, which are coupled to the nasal tract.

The reconstruction of a continuous fundamental phase for speech segments with uninterrupted phonation opens the possibility to complement the analysis of the spectral properties of the speech signal by a run time analysis. The run time differences may refer either to different paths of the response to the early (sonic) acoustic excitation, which is created in the vicinity of the glottis, or to different speeds of the action of the fundamental drive on the retarded (subsonic) acoustic excitation, which is created in the vicinity of a phoneme specific constriction of the vocal tract. The delay of the retarded action results from subsonic transport of convective energy to the site of the enhanced production of acoustic energy. The delay depends in particular on the relative share of subsonic transport on the way from the glottis to the secondary site of acoustic excitation. In the case of the fricatives the high precision determination of the delay cannot only be achieved by parameter estimation of the delay τ in equation (6) but also by inspection of the fundamental phase profile of the primary (or secondary) response. As has been demonstrated by Jackson and Shadle [19] fricatives show characteristic delays of the amplitude (envelope) maximum of the subsonic (unharmonic) excitation. Both types of run time differences are potentially suited as additional cues for phoneme recognition.

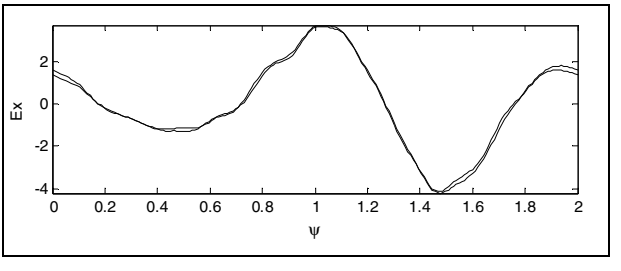

Fig. 5. Fundamental phase dependent coupling function $G_{j,2}(\psi)$ reconstructed with periodicity $p = 2$ for the vowel of the first occurrence of the word "wind" used in figure 3. The two curves correspond to the odd and even periods. The good agreement can be interpreted as a hint to the high robustness of the reconstruction of the excitation.

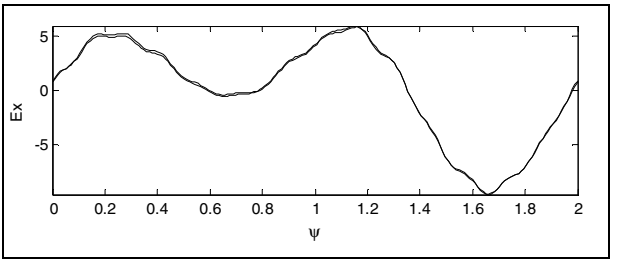

Fig. 6. Fundamental phase dependent coupling function $G_{j,2}(\psi)$ for the nasal of the word "wind" of figure 5

In the case of the higher frequency subbands of voiced fricatives the interference between the responses of the sonic and the subsonic excitation may lead to a sensitive dependence on the recent history of the fundamental phase. This mechanism of deterministic amplification of the aperiodicity of the fundamental drive process is suited to explain (a part of) the typical aperiodicity of the higher frequency subbands of sustainable voiced fricatives. The perceptional relevance of the "drive – response chaos" of voiced fricatives should be analysed by appropriate psycho-acoustic experiments. The described two-level drive – response model is well suited to support such experiments.

5 Discussion and Conclusion

The present study is based on the link between speech-physiology and psycho-acoustics, which results from the phylogenetic and ontogenetic coevolution of the auditive pathway and the sound production system in an acoustic environment, which is strongly influenced by sound utterances of contemporary members of the own species. The assumption that the far field acoustic response of the pulsed turbulent airflow in the vocal tract can be described by a low dimensional synchronisation manifold, is a remarkable hypothesis, which should be interpreted as result of the ontogenetic adaptation of human speech production. The success of the proposed description of non-pathological voiced speech relies to a large extent on the precision of the reconstructed fundamental phase. The robustness and generality of the present method to extract the fundamental phase out of a speech signal is not yet comparable to the one of human pitch perception. However, the newly established link between speech-acoustics and psycho-acoustics can be exploited as a guide to future improvements of the reconstruction of the fundamental drive.

The transmission protocol of voiced human speech is based on the production and analysis of complex airflow pattern in the vocal tract of the transmitter. The present study demonstrates that the analysis on the receiver side can be focussed on the mode locking of the pulsed airflow by replacing the time dependent excitation of the classical source - filter model by a fundamental phase dependence which can be described by a low dimensional generalized synchronization manifold. In the simpler cases of vowels and nasals the manifold (surface) can be described by a single coupling function, which depends on a single fundamental phase. In the case of voiced consonants with a phoneme specific constriction of the vocal tract the excitation may have to be extended or replaced by a coupling function, which depends on a delayed fundamental phase. The evolution of speech has lead to many voiced phonemes, which can be distinguished by properties of these coupling functions and the closely related two-level drive - response models. To make the coupling functions visible (or audible) with increased precision, a voice specific subband decomposition of the speech signal has been proposed, which is suited to extract the phase of the fundamental drive with high precision. The extraction relies on the fact that non-pathological voiced speech leaves several subbands undistorted by vocal tract resonance or phoneme specific constriction of the airflow.

There have been numerous attempts to increase the precision of the spectral analysis of voiced speech by introducing a "dynamic time warping" preprocessing step [33] which enhances the proportionality between the (artificial) time and the fundamental phase. Such a preprocessing step, however, ignores the dynamic nature of the

production of voiced speech which involves phoneme specific delays of the primary response and a fully dynamic secondary response. The dynamics of the secondary response may show a sensitive resonance behaviour with respect to changes in the time scale. A time warping of the speech signal, which enhances the visibility of the synchronization manifold of the primary excitation, can thus be expected to have a non-negligible corrupting effect on the spectrum of the secondary response.

In the case of vowel – nasal and vowel – fricative transitions, in particular, the response of the fundamental drive may show a long range correlation with a delay which exceeds the length of the conventional window of analysis. In these cases a phase vocoder, which is based on a continuously reconstructed fundamental drive process and a related two level drive – response model with appropriate time delays, is expected to solve some of the major coarticulation problems of present day phase vocoders, which, so far, have prevented them to replace concatenative synthesis in high quality speech reconstruction.

Acknowledgement. The author would like to thank V. Hohmann, B. Kollmeier and J. Nix, Oldenburg, M. Kob, C. Neuschaefer-Rube, Aachen, G. Langner, Darmstadt, N. Stollenwerk, Porto, J. Schoentgen, Brussels, J. Rouat, Montreal, P. Grassberger, M. Schiek and P. Tass, Jülich for helpful discussions.

References

[1] Fant G. *Acoustic theory of speech production*, Mouton, 'S-Gravenhage (1960)
[2] Vary P., U. Heute, W. Hess, *Digitale Sprachsignalverarbeitung*, B.G. Teubner Verlag, Stuttgart (1998)
[3] Schroeder M.R., *Computer Speech*, Springer (1999)
[4] Titze I.R., *Acta Acustica* **90**, 641-648 (2004)
[5] Kantz H., T. Schreiber, *Nonlinear time series analysis*, Cambridge Univ. Press (1997)
[6] Kocarev L., U. Parlitz, *Phys. Rev. Lett*. **76**, 1816 (1996)
[7] Schoentgen J., "Stochastic models of jitter", *J. Acoust. Soc. Am.* **109** (4): 1631-1650 (2001)
[8] Herzel H., D. Berry, I.R. Titze and I. Steinecke, „Nonlinear dynamics of the voice: Signal analysis and biomechanical modeling ", *Chaos* **5**, 30-34 (1995)
[9] Seebeck A., "Über die Sirene", *Annalen der Physik*, LX, 449 ff, ibid. LXIII, 353 ff and 368 ff (1843)
[10] Moore B.C.J., *An introduction to the psychology of hearing*, Academic Press (1989)
[11] De Cheveigné A. and H. Kawahara, „Comparative evaluation of F0 estimation algorithms", Eurospeech 2001, Alborg (2001)
[12] Winholtz W.S. and L.O. Ramig, "Vocal tremor analysis with the vocal demodulator", *J.Speech Hear. Res.* **35**, 562-573 (1992)
[13] Hanquinet J., F. Grenez and J. Schoentgen, "Synthesis of disordered voices", *this volume* (2005)
[14] Kubin G., "Nonlinear processing of speech," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 557–610, Amsterdam: Elsevier (1995)
[15] Moakes P. A. and S.W. Beet, "Analysis of non-linear speech generating dynamics," in *ICSLP 94,* Yokohama, pp. 1039–1042 (1994)
[16] Drepper F.R. in C. Manfredi (editor), *MAVEBA 2003*, Firenze University Press (2004)
[17] Drepper F.R., „Selfconsistent time scale separation of instationary speech signals*", Fortschritte der Akustik-DAGA'05* (2005)

[18] Teager H.M. and S.M. Teager, “Evidence for nonlinear sound production mechanisms in the vocal tract,” in *Proc NATO ASI on Speech Production and Speech Modelling*, pp. 241–261 (1990)
[19] Jackson P.J.B. and C.H. Shadle, “Pitch scaled estimation of simultaneous voiced and turbulence-noise components in speech”, *IEEE trans. speech audio process.,* vol. **9**, pp. 713-726 (2001)
[20] Maragos P., J.F. Kaiser and T.F. Quatieri, “Energy separation in signal modulations with application to speech analysis", IEEE Trans. Signal Processing, Vol. 41, pp. 3024-3051 (1993).
[21] Zhao, W., C. Zhang, S.H. Frankel and L. Mongeau, “Computational Aeroacoustics of Phonation, Part I: ... ”, *J. Acoust. Soc. Am.*, Vol. **112**, No. 5, pp. 2134-2154 (2002)
[22] Hohmann V., *Acta Acustica* **10**, 433-442 (2002)
[23] Zwicker E. und Feldtkeller R., *Das Ohr als Nachrichtenempfänger*, Hirzel Verlag, (1967)
[24] Sottek R., *Modelle zur Signalverarbeitung im menschlichen Gehör*, Verlag M. Wehle, Witterschlick/Bonn (1993)
[25] ftp.cs.keele.ac.uk/pub/pitch
[26] Drepper F.R., “Rekonstruktion stationärer Mannigfaltigkeiten der Teilbanddynamik instationärer Sprachsignale” *Fortschritte der Akustik-DAGA’03* (2003)
[27] Drepper F.R., „Voiced excitation as entrained primary response of a reconstructed glottal master oscillator*”, Fortschritte der Akustik-DAGA’05* (2005)
[28] Afraimovich V.S., N.N. Verichev, M.I. Rabinovich, *Radiophys. Quantum Electron*. **29**, 795 ff (1986)
[29] Rulkov N.F. , M.M. Sushchik, L.S. Tsimring, H.D.I. Abarbanel, *Phys. Rev. E* **51**, 980-994 (1995)
[30] Rulkov N.F. , V.S. Afraimovich, C.T. Lewis, J.R.Chazottes and A. Cordonet, *Phys. Rev. E* **64**, 016217 (2001)
[31] Schoentgen J., “Shaping function models of the phonatory excitation signal”, J. Acoust. Soc. Am. **114** (5): 2906-2912 (2003)
[32] Kawahara H. and P. Zolfaghari, “Systematic F0 glitches around nasal-vowel transitions”, *Eurospeech 2001* (2001)
[33] Graf J. T. and N. Hubing, “Dynamic time warping comb filter for the enhancement of speech degraded by white Gaussian noise,” *Proc. ICASSP*, vol. **2**, pp. 339–342, (1993)