Abstract: Voiced speech is characterized by qualitatively rich mode locking phenomena linking harmonically excited acoustic modes of the vocal tract. Due to the strong instationarity of speech, a differentiated analysis of these modes cannot be achieved with the help of a linear, time invariant source and filter model (based on stationary sources). As alternative, the characteristic mode locking is described as generalized synchronization in drive - response systems with an instationary, common (fundamental) drive. By introducing a combined harmonic and logarithmic (audio-logical) scale subband decomposition adapted to the frequency of the master oscillator of phonation, a self-consistently confirmed, topologically equivalent reconstruction of a number of acoustic modes of an acoustic object is generated. Whereas the invariant resonator properties (Lyapunov exponents) of the reconstructed response dynamics are characteristic for vowels, the generalized synchronization manifolds (lines or surfaces) in the combined state space of drive and response band can be used for the distinction of consonants. The topologically equivalent reconstruction of the phonation process is potentially useful for phoniatric diagnoses.

Keywords: Subband decomposition, drive – response reconstruction, transfer function model, voiced speech, generalized synchronization

I. INTRODUCTION

The characteristic mode locking of voiced speech results from harmonic excitations, which are synchronized by glottal closure events [1]. In the context of generalized synchronization in drive – response systems it has been shown recently [2,3], that mode locking or synchronization is not an elementary phenomenon, but a header for a larger number of qualitatively different coordination possibilities, which are characterized by more or less smooth and/or continuous invariant manifolds in the combined state space of coupled drive - response oscillator pairs, the manifolds being defined by maps, which relate a state of the response uniquely to the simultaneous state of the drive [2-4]. In the context of speech recognition the topological equivalence between drive and response represents an important special case [5], which is characterized by a conjugation (a continuous and uniquely invertible map). Together with the more general concept of conditional asymptotic stability [6] these notions are useful for a differentiated analysis of the synchronization or coordination phenomena of voiced speech.

The ubiquitous instationarity of the amplitude and pitch of phonation is a second essential feature of voiced speech, the variation of the amplitude being relevant on time scales down to less than 50 ms. In this context it is important to note that the long time known phenomenon of synchronization is not limited to periodic or quasi periodic driving but may as well occur for stochastic [7] or deterministic chaotic driving [2]. So far the application of the source and filter model to the recognition of voiced speech is based on the assumption of a stationary phonation process [8]. This assumption limits the source and filter model to the description of relatively short sections of speech (typically 20 ms). Such short sections, however, are insufficiently suited to detect the characteristic invariant manifolds of voiced phonemes. The ubiquitous instationarity of human speech motivates, to replace the assumption of stationary phonation (implicitly implied when estimating spectra) by the assumption of generalized synchronization between the instationary and/or nonlinear drive and the acoustic response. Thus the atoms or objects (in particular the phonemes) of speech are no longer interpreted as stationary processes but as stationary or invariant manifolds (lines or surfaces) in the combined state space of instationary drive and response oscillator pairs. However, neither the acoustic response modes within the vocal tract nor the excitation within the glottis can directly be observed in the situation of speech communication.

II. SUBBAND RECONSTRUCTION

As a characteristic feature the present approach uses suitably chosen bandpass filters to determine a fundamental driver mode as well as higher frequency subbands, which represent topologically equivalent reconstructions of corresponding acoustic modes of the vocal tract. The choice of the appropriate bandpass filters is based on the fact that voiced speech is characterized by a concentration of power in comparatively narrow frequency ranges and that due to the approximate periodicity of the voice source these frequency ranges show a comb like pattern, aligned to the fundamental frequency defined as (short time) average of the frequency of glottal closure events. The bandwidths of the bandpass filters should be chosen
sufficiently narrow, to resolve as many harmonics as possible, however also sufficiently broad, such that the relative bandwidth exceeds the one of the instationary frequency fluctuations of the fundamental drive process. Obviously the ERB bandwidths (according to the equivalent rectangular bandwidth model) [9,10], known from masking experiments in psychoacoustics, represent an evolutionarily successful compromise. This choice introduces an a priori limit on the harmonic number $h$ of resolvable subbands, ($h < 10$). When generating a vowel, the vocal tract shows no branching and no additional constriction (apart from the glottis). In this situation the feasibility of a harmonic scale aligned subband decomposition is guaranteed, since the response processes of the different harmonic excitations superepose without perturbation and can thus be separated by appropriate bandpass filters due to their differing frequencies.

Even in the case of nasals or voiced approximants like /l/ or /v/ in veal and voiced sibilants like /th/ in thumb, the concentration of power of the primary voice source (in space and frequency) implies or supports a phonation dynamics, which features a causal pinhole expressed by a low dimensional, potentially instationary master oscillator, which “enslaves” [4] the faster state variables of sound production or at least their long distance effect on the acoustic field. According to the so far rather limited study there is no contradiction, that at least in the case of healthy phonation the voiced part of the excitation of the acoustic modes can be expressed as synchronization manifolds, which are driven by a pair of fundamental amplitude and phase. The complex wavelet transformation [11,12],

$$A_{j} e^{i \psi_{j}} = \sum_{t} X_{tk} (e^{j0.5t/a} - e^{-0.5t/a^2}) e^{-0.5k^2/a^2},$$

turns out to be particularly suited for the extraction of the amplitude $A_{j}$ and phase $\psi_{j}$ of the master oscillator from the speech signal. The centre frequency $\omega_{c}$ is chosen as an appropriate multiple of the fundamental frequency $F_{0}$, which is obtained by a conventional method.

Following the well accepted linear source and filter model of speech production [1,8] it is plausible to represent the voiced part of each subband specific excitation as product of drive amplitude $A_{j}$ and an oscillatory driver phase dependent excitation function $G_{j}(\psi_{j})$, which thus takes the central role in the phenomenological description of complex voiced phones. The enslavement of the fast degrees of freedom of the excitation implies a periodicity of the excitation function. In the context of instationary phonation it is important to note that this periodicity does not refer to time but to the phase of the glottal drive. The period length $2\pi p_{j}$ of the excitation function is potentially speaker dependent and coincides usually with the fundamental period $2\pi$. Due to the band limitation each excitation function can nicely be approximated by a finite Fourier series, the terms of which may be interpreted as purely harmonic, elementary excitations.

Following the linear source and filter model, subband $j \{X_{jt} | t = 0, 1, \ldots \}$ is approximated as a finite dimensional, linear response to a drive synchronous excitation. Due to the described band limitation of the subbands as well as due to the band adapted time step length $\Delta$ (chosen as a quarter of the period length defined by the band specific central filter frequency) a two dimensional response dynamics turns out to be sufficient,

$$X_{j,(n+1)\Delta} = a_{j} X_{j,n\Delta} + b_{j} X_{j,(n-1)\Delta} + A_{n\Delta} G_{j}(\psi_{n\Delta})$$

with $G_{j}(\psi_{n\Delta}) = \sum_{k=0}^{K_{n} \Delta} c_{j,k} \cos(k\psi_{n\Delta} / p_{j} - \gamma_{j,k})$,

$$n = 0, 1, \ldots \text{ and } K_{n} \Delta \leq 2 h_{j}, \text{ where } h_{j} \text{ represents the band index dependent harmonic. The goal of the phonation process adapted bandpass decomposition is characterized by subbands, which can be approximated as two-dimensional response to a single, pure harmonic, elementary excitation. In the case of the higher harmonic subbands, in particular of consonants, the goal reduces to maximal diagonal dominance of the subband specific elementary harmonic excitation. The average distance of index $k$ to the band specific harmonic $h_{j}$ turns out to be a useful objective function,

$$\bar{\Delta k}_{j} = \frac{1}{K_{n} \Delta} \sum_{k=0}^{K_{n} \Delta} c_{j,k}^{2} |k - h_{j}|.$$.

The central filter frequency of the fundamental subband filter represents the essential adaptation parameter to achieve the diagonal dominance of the elementary excitations.

### III. Topological Equivalence

The introduction of time dependent and time related (continuously extended, unwrapped) phases as state variables of the response dynamics opens the possibility to identify (1:n) or (n:m) mode- or phase locking as a near linear, diffeomorphic conjugation. Due to transitivity and invertibility of conjugations in a chain of conjugated oscillators, the evidence of a near linear conjugation between the subband oscillators of a voiced signal can be taken as a confirmation of the topological equivalence of all oscillators involved, including the equivalence between the respective harmonically excited acoustic mode within the vocal tract and the corresponding subband (figure 1). The confirmation of topological
Voice analysis 279

Invariants [13] of the resonator dynamics. According to the so far rather limited empirical basis (50 ms subsections of 5 vowels and 6 sustainable voiced consonants uttered by 4 male and 2 female subjects), the described subband decomposition of voiced phonemes generally offers the possibility to detect near linear conjugation between the lower harmonic subbands (figure 1). This way the phase and amplitude of the fundamental drive can generally be confirmed as topologically equivalent image of state variables of the fundamental glottal mode. The presented approach is thus well suited for a robust and precise determination of the momentary pitch of voiced speech and potentially also for phoniatric diagnoses.

For subbands within the harmonically resolvable range (harmonic number \( h < 9 \)), a missing conjugation to the driver band can be attributed to a break up of the conjugation chain within the vocal tract and not to a break up on the way from the vocal tract to the ear or microphone (figure 1). In the case of voiced approximants and sibilants, in particular, the loss of conjugation between subbands does not indicate a loss of conditional asymptotic stability [6] of the higher harmonic subbands. A general definition of complex voiced phones of human speech can thus be given as existence of a bandpass filter based subband decomposition, which contains one fundamental drive oscillator and further conditionally stable response bands, where the conditioning is limited to the amplitude and phase of the drive and where the drive can be confirmed to be (1:1) equivalent to the fundamental glottal oscillator.

Strikingly many distinctive properties of voiced phonemes coincide with topological invariants of the response dynamics or with topologically invariant geometric properties of the related invariant manifolds. The most important topological invariants of the subband dynamics are the (conditional) Lyapunov exponents [6], since they express resonator properties of the vocal tract, like resonator quality and eigen-frequency, which are known to be strongly dependent on the geometry of the vocal tract and thus particularly suited for the distinction of vowels. The distinctive properties of consonants are predominantly related to geometric properties of invariant manifolds in the four-dimensional state space of drive - response oscillator pairs (like kinks or jumps in the case of nasals). Stop consonants are characterized by a pronounced visibility (audibility) of the amplitude – amplitude coupling between the drive and the respective response bands, whereas for sustainable voiced consonants the coupling of the response phases to the driver phase plays the more important distinctive role.

IV. EVOLUTIONARY ASPECTS OF VOICED SPEECH

As a striking feature of human speech, the confirmation of the topological equivalence can often be achieved for subbands with harmonic numbers higher than 10. (Due to resonant excitation, the detectability of higher harmonic, phase locked modes becomes extreme in the case of singing.) The surprisingly extended success of the described approach towards the determination of phonation and vocal tract equivalent excitation and response processes, can only be explained within the framework of evolutionary and ontogenetic adaptation, characterized by a near optimal fit between properties of human speech and the abilities of auditive perception. Thus voiced speech and singing have to be interpreted as results of adaptation processes, which favor easy detectability within a confusion of voices.

In view of the pronounced differentiation of the synchronisation phenomena of voiced speech, auditive perception of humans can be assumed to be able to perform and select the skilled bandpass decomposition, which uncovers the more or less smooth, stationary manifolds in the combined state space of the subbands - even in the case of instationary phonation. There are several empirical facts, which support a perception equivalent model of hearing, which is build on the described synchronisation analysis of voiced speech. Firstly there is the central role of the pitch known to be relevant on different semantic layers of speech communication and to be perceived even in the case of imperfect harmonicity [14]. Further support can be seen in the astonishing monaural voice separation and speaker identification ability of the auditive perception of humans, which (in particular in the case of rough phonation) could so far neither be explained by perceptual models nor imitated by speech - or speaker recognition algorithms.

Based on highly developed abilities of higher vertebrates [15,16], the astonishing speaker identification ability indicates that the auditive perception of humans is in command of analysing abilities of the nonlinear dynamics of phonation, including recognition of subharmonics or of co-existing meta-stable periodic trajectories (unstable periodic orbits, UPO’s) [3,17]. In order to avoid dangerously large bandwidths of the fundamental driver mode it is advantageous to represent the influence of the mentioned nonlinear phonation dynamics with the help of periodicity \( p_j \) of the driver phase dependent excitation function. The potential richness of the combination possibilities of periodicity \( p_j \) of an excitation manifold with the periodicity \( q_j \) of the resulting response manifold and the winding number \( w_j \) of the corresponding response phase offers a plausible explanation for the astonishing speaker recognition ability of auditive perception.
V. CONCLUSION

Contrary to the conventional approach towards speech analysis, which is based on the assumption of a stationary, high dimensional source and the use of a broadband version of the linear source and filter model, the newly proposed approach describes the source as a synchronized response to a low-dimensional instationary drive, which is determined self-consistently as a topologically equivalent image of the underlying fundamental glottal mode. The self-consistency is based on a skilled subband decomposition, the subbands of which can optimally be interpreted as linear response to harmonically distinct, voiced excitations. Apart from providing evidence of the topological equivalence of the common drive, the skilled subband decomposition discloses topologically equivalent images of the invariant manifolds, which characterize the synchronization of the higher harmonic acoustic modes of the vocal tract. The distinction of consonants is hypothesized to rely largely on topologically invariant geometric properties of these manifolds. Since the parameters of the excitation manifolds can be estimated efficiently with the help of multiple linear regression, the outlined synchronization based analysis of voiced speech is expected to be feasible in real time.


REFERENCES


Figure 1: Voiced phones of human speech are characterized by stationary manifolds (lines or surfaces) in the combined state space of drive and response oscillator pairs, which differ with respect to the distance to sound generation inside the glottis as well as with respect to the respective oscillation or winding number \( h \) of the subband specific excitation. The dynamics of the excitations as well as of their resulting response processes are reduced to the corresponding phase dynamics, the respective driver phases \( \Psi \) or \( \Psi ' \) being indicated horizontally and the corresponding response phases vertically.