Voiced excitation as entrained primary response of a reconstructed glottal master oscillator

F.R. Drepper
Forschungszentrum Jülich GmbH, D 52452 Jülich, Germany
f.drepper@fz-juelich.de

The transmission protocol of sustained 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 a 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 synchronization manifold described by speaker and phoneme specific coupling functions. In the case of vowels and nasals the excitation depends on a single phase of the fundamental drive. In the case of other sustained voiced consonants the excitation may include an additional coupling function, which depends 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 by methods assuming stationary excitation.

Keywords: voiced speech, fundamental drive, two-level drive-response model, generalized synchronization, delayed excitation

I. INTRODUCTION

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 secondary 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.

Due to the strong nonlinearities of the coupled dynamics, non-pathological, standard register phonation dynamics is characterized by a stable synchronization 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 non-singular invertible mapping (conjugation) [5]. 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. Therefore a glottal master oscillator can be defined, which enslaves (synchronizes) or drives the other oscillators including the higher frequency acoustic modes.

In the case of non-pathological voiced speech the observation of the air pressure signal or of the electro-glottogram reveals a unique frequency of phonation, the fundamental frequency. 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 [6, 7]. Except at the high frequency end the deviation of the glottal cycle length 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 [6, 7]. As a general feature, cycle length differences increase with the time scale (the relative differences ranging from less than 1 % up to more than 20%). In spite of the partially minor amplitudes of aperiodicity all or most of these frequency bands appear to be perceptually relevant. 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 mainstream approach of speech analysis this has lead to the more or less explicit assumption that the excitation is wide sense stationary in the analysis window, which is usually chosen as 20 ms [2, 3]. The latter assumption 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 [9, 10]. 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 [9, 10]. The assumption of a low dimensional dynamical system, however, is in contradiction to the observed complexity 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 [11-12]. 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 [11]. 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 [11-12].

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 [14] pointed out that the conversion of the potential energy of the compressed air in the subglottal
and with a subband inde-
lead to a
0.52
with .         (2) 1
0.53
j
1
0.51
is based on an adap-
with a high time and frequency resolution. Based on a,
[18] and 1
are given in a partially unwrapped form, depen-
j
0.51
as part of a two-level drive-response model, which extends the
validity range of the classical source-filter model and which is suited
fundamental drive. In the case of a subband with paired harmonics,
result from (n:1) and (m:1) phase relations to the fundamental drive.
reconstruction of the glottal master oscillator for segments of
consonants, which are extremely difficult to analyse in vivo [14], in
vivo [15] and in silico [15].

II. 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 4th order complex gammatone bandpass filters with roughly
approximate audiological bandwidths $\Delta F_j$ and with a subband inde-
pendent analysis - synthesis delay as described in Hohmann [16].

The extraction of the fundamental phase $\varphi_j$ is based on an adap-
tation of the best filter frequencies $F_j$ of the subband decomposition to the momentary frequency of the glottal master oscillator (and its higher harmonics). 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.

$$F_j = \left\{ \begin{array}{ll}
\frac{j F_1}{\sqrt{j(j+1) F_1}} & \text{for } 1 \leq j \leq 6 \\
\frac{F_1}{2 F_1} & \text{for } 6 < j \leq 11
\end{array} \right. \quad (1a)$$

$$\Delta F_j = \left\{ \begin{array}{ll}
\frac{F_1}{2 F_1} & \text{for } 1 \leq j \leq 6 \\
\frac{F_1}{6 F_1} & \text{for } 6 < j \leq 11
\end{array} \right. \quad (1b)$$

It is further assumed that voiced sections of speech are produced with at least two subbands, which are not distorted by vocal tract resonances or additional constrictions of the airflow. In the case of subbands with separated harmonics, $1 \leq j \leq 6$, 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, $6 < j \leq 11$, 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 sections of speech the iterative improvement leads to a fast converging fundamental phase velocity $\varphi_j$ with a high time and frequency resolution. Based on a, so far, arbitrary initial phase, successive estimates of $\varphi_j$ lead to a reconstruction of the fundamental phase $\varphi_j$, which is uniquely defined for uninterrupted segments of voiced phonation.

The fundamental amplitude $A_j$ is assumed to be related to loudness perception [17] by a power law. The exponent $1/\nu$ is chosen such that the fundamental amplitude represents a linear homogenous function of the time averaged amplitudes $\overline{A}_j$ of a synthesis suited set of subbands,

$$A_j = \left( \sum_{j=1}^{N} (g_j \overline{A}_j) \right)^{1/\nu} \quad \text{with} \quad \sum_{j=1}^{N} g_j^\nu = 1. \quad (2)$$

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 $g_j = h_j^\nu$, where $h_j$ represents the (integer) centre harmonic number, which approximates the ratio $F_j / F_1$. The present study uses $\nu = 0.3$ [18] and $\mu = 1$ [3]. The synthesis suited set of subbands is generated by replacing the over complete subband set $6 < j \leq 11$ by a set $6 < j \leq N$, which is spaced equidistantly on the logarithmic frequency scale with 4 filters per octave,

$$F_j = 5 \cdot 2^{(j-6)/4} F_1 \quad \Delta F_j = 2^{(j-5)/4} F_1 \quad (3)$$

The feasibility of the extraction of the fundamental drive as well as the validity of its interpretation 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 [19]. 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. 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 $\Phi_j$ are given in a partially unwrapped form, depend-
ing on the respective centre harmonic number $h_j$. The enlarged

![Figure 1](image_url)

**Figure 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 [19]. The lower panel shows the reconstruction of the fundamental phase $\varphi_j$ in units of $\pi$. The time scale (in units of seconds) corresponds to the original one.
system describes the more or less resonant “signal forming” on the way through the vocal tract as action of a linear autoregressive filter. The subband decomposition (1) and (3) being used for the reconstruction of the fundamental drive can also be used with advantage to achieve a numerically robust reconstruction of the excitation.

III. ENTRAINMENT OF THE PRIMARY RESPONSE

Due to the slow velocity of the glottal tissue (compared to the velocity of sound) the excitation $E_{j,p}$ of a voiced subband with index $1 \leq j \leq N$ can be assumed to be restricted (enslaved or entrained) to a generalized synchronization manifold (surface) in the combined state space of drive and response [20-22]. In the simplest case the time dependence of subband excitation $E_{j,p}$ can thus be replaced by a dependence on the simultaneous state of the fundamental drive. More generally, the dependence of the state of the primary response on the state of the fundamental drive may degenerate to a multi-valued mapping, which can, however, be expressed by a unique function of the unwrapped fundamental phase $\psi_j$ [11-12].

$$E_{j,p} = A_j \ G_{j,p}(\psi_j) = A_j \sum_{k \in \mathbb{Z}} c_{j,k} \ \exp(ik \ \psi_j). \quad (4)$$

As part of the improved time scale separation the generalized synchronization manifold is assumed to be the product of the slowly variable fundamental amplitude $A_j$ and the potentially fast varying complex coupling function $G_{j,p}(\psi_j)$, the real part of which describes the subband excitation. In its general form, $G_{j,p}(\psi_j)$ represents a $2\pi \ p$ periodic function of the unwrapped fundamental phase $\psi_j$ with an integer period number $p \geq 1$ and can thus be well approximated by the finite Fourier series in equation (4). Voiced excitations are characterized by 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 [20], whereas $p = 2$ is suited to describe the period doubling voice type [4]. The unwrapped fundamental phase can be assumed to be approximately proportional to time. When $2\pi \ p$ exceeds the length of the analysis window, equation (4) is therefore suited to describe a fully general excitation, including the unvoiced case.

The excitation parameters $c_{j,k}$ 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. To achieve a comparable numerical robustness, the parameter estimation is done separately for the different frequency bands. The band limitation can be used to reduce the number of resonances (poles of the autoregressive filter), which are relevant for the respective subband. The complex subband $\{X_{j,k}\}$ can thus be described by the following nonlinear conditional stochastic process with a two-level drive – response model as deterministic part (skeleton) [11-12],

$$X_{j,k+1} = b_j \ X_{j,k} + A_j \ G_{j,p}(\psi_j) + A_j \ \sigma_j \ \xi_{j,k}. \quad (5)$$

where $\Delta$ denotes the subband specific prediction step length, $b_j$ the complex subband specific resonator parameter, $\xi_{j,k}$ a (0,1) Gaussian complex white noise process and $\sigma_j$ the time independent part of the standard deviation. 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 summation index set $S_{j,k}$ of equation (4) is chosen in accordance to the respective bandpass filter. The decomposition into subbands is used to estimate equation (5) with a subband specific integer time step...
length $\Delta$. The aggregated coupling function, which results from the sum of all subband specific coupling functions, can be compared to the excitation of the (single level) broadband source - filter model.

Vowels and nasals are characterized by the fact that the time points of the 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.

In the case of voiced speech segments, which contain other sustainable voiced consonants, the continuous reconstruction of the fundamental phase can be used advantageously to extend equation (5) by a second excitation term $A_2 G_2(p_{2\omega})$ with a coupling function, which depends on a delayed fundamental phase. According to Teager and Teager [14] the delay $\tau$ can be interpreted as result of the comparatively slow subsonic convective transport of kinetic energy to the site of the phoneme specific secondary constriction of the vocal tract, where the conversion to acoustic energy takes place.

IV. DETERMINISTIC APERIODICITY AMPLIFICATION OF VOICED CONSONANTS

As a striking result, the assumption of generalized synchronization of the primary response does not only hold in the case of vowels but also in the case of many sustained voiced consonants. In the case of the voiced approximant /l/ the aggregated coupling function shows several steep slopes which indicate a sensitive dependence on the phase of the fundamental drive (figure 4). The sensitive dependence can be interpreted as effect of the superposition of the response of the direct excitation and the one of the delayed excitation resulting from an intermittently turbulent airflow. The interference between the two responses may lead to a sensitive dependence on the recent history of the fundamental phase. First results show that the deterministic aperiodicity amplification is a widespread feature of voiced speech. Its occurrence shows a marked dependence on the speaker and on the fundamental phase.

![Figure 4: Aggregated fundamental phase dependent coupling function reconstructed with period $p = 2$ for the voiced approximant /l/ of the word “along” uttered by the first male speaker. The two curves correspond to the odd and even periods. The disagreement of the two curves shows a marked dependence on the fundamental phase.](image)

The continuous reconstruction of the 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 a travel time difference of the primary acoustic pulse or to a build up time of the turbulence at the secondary constriction of the vocal tract. The coupling functions with periodicity $p = 2$ are suited to describe a voice type, which cannot be classified uniquely by using cycle lengths differences (figure 4). It is hypothesized that the fundamental phase dependent coupling functions are suited to serve as additional cue for phoneme recognition and as fingerprint for speaker identification.

V. CONCLUSION

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 dependence of the excitation of the classical source - filter model by a fundamental phase dependence, which can be described by a low dimensional generalized synchronization manifold (surface or coupling function). The evolution of speech has lead to many voiced phonemes and syllables which can be distinguished by properties of one dimensional coupling functions and of a closely related two-level drive - response model. To make the coupling functions visible with increased precision, a voice specific subband decomposition of the speech signal has been proposed, which is suited to extract a precise fundamental phase. The extraction relies on the fact that non-pathological voiced speech leaves at least two subbands undistorted by vocal tract resonance or secondary constriction.


References

[18] ftp.cs.keele.ac.uk/pub/pitch