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

F.R. Drepper

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

A time scale separation of voiced speech signals is introduced, which avoids the assumption of a frequency gap between the acoustic response and the prosodic drive. The non-stationary drive is extracted selfconsistently from a voice specific subband decomposition of the speech signal. When the band limited prosodic drive is used as fundamental drive of a two-level drive-response model, the voiced excitation can be reconstructed as a trajectory on a generalized synchronization manifold, which is suited to serve as cue for phoneme recognition and as fingerprint for speaker recognition.

1. Introduction

The vocal tract excitation of voiced speech is generated by a pulsatile airflow, which is strongly coupled to the dynamics of the vocal fold. The voiced excitation is created in the vicinity of the vocal fold and/or as acoustic (high frequency) part of an intermittently turbulent airflow in the vicinity of a secondary constriction of the vocal tract [1]. 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 a glottal master oscillator can be defined, which plays the role of an autonomous drive of the vocal tract excitation.

Time series of successive cycle lengths of the glottal 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 frequency end the cycle lengths represent a non-stationary stochastic process. Several frequency bands, including the higher frequency ones, are known to play a major role for the nonsymbolic information content of speech. The relevant frequency range of the vocal tract excitation extends about two orders of magnitude higher than the fundamental (glottal) frequency. It is therefore common practice to introduce a time scale separation, which separates the high frequency acoustic phenomena of speech signals from the frequency range of the subharmonic, subacoustic and prosodic aperiodicity of the glottal oscillator [2, 3].

A simple approach to time scale separation is to assume a frequency gap, which separates the autonomous lower frequency degrees of freedom from the higher frequency (dependent) ones. In the case of speech analysis this has lead to the more or less explicit assumption that the acoustic excitation of the speech signal is wide sense stationary in the analysis window, which is usually chosen as 20 ms [2, 3]. A comparatively well elaborated attempt to avoid this assumption is given in Kawahara et al. [4]. Whereas the simplifying assumptions of the present approach are focussed on the synchronization analysis of the pulsatile airflow, the assumptions of [4] are focussed on a flexible speech synthesis.

The present study avoids the assumption of a causal frequency gap by treating the broadband excitation as a stationary response of a non-stationary, band limited fundamental drive, which is extracted selfconsistently from voiced sections of the speech signal [5-7]. The selfconsistency refers to the confirmation that the fundamental drive can be interpreted as a topologically equivalent reconstruction of the glottal master oscillator, which synchronizes the voiced excitation [5, 6]. To analyse synchronization or mode locking phenomena between non-stationary subsystems it is useful to determine the phases of band-limited oscillators [8]. In contrast to the prevailing speech analysis, which considers time independent phases of Fourier components with zero bandwidth, the present approach is based on time dependent phases of subbands with finite bandwidths. To determine the latter type of phases it is useful to describe all subband or oscillator states by complex variables.

2. Extraction of the fundamental drive

As an important property of non-pathological voiced human 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. The (response related) state of the fundamental drive should not be confused with the state of a dynamical system suitable to describe the self-sustained oscillations of the glottis. As has been pointed out by Titze [9], a mechanistic model of such a system cannot be restricted to state variables of the vocal fold alone. The phase of the glottal master oscillator should rather be interpreted as a phase, which uniquely describes a state on the limit cycle, which attracts non-pathological, self-sustained oscillations of the glottis. As a characteristic feature of the present phenomenological approach, the amplitude and phase of the fundamental drive are extracted from subband decompositions of the speech signal. The decompositions use 4^{th} order complex gammatone bandpass filters with roughly audiological bandwidths ΔF and with a subband independent analysis - synthesis delay as described in Hohmann [10].

The extraction of the fundamental phase ψ_t is based on an adaptation 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 centered 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 centered on pairs of neighbouring harmonics.

$$F_j = \left\{ \begin{array}{c} j\,F_1 \\ \sqrt{j\,(j+1)}\;F_1 \end{array} \right\} \quad \text{for} \quad \left\{ \begin{array}{l} 1 \le j \le 6 \\ 6 < j \le 11 \end{array} \right\} \qquad (1a)$$

$$\Delta F_j = \left\{ \begin{array}{c} F_1 \\ 2\,F_1 \end{array} \right\} \quad \text{for} \quad \left\{ \begin{array}{l} 1 \le j \le 6 \\ 6 < j \le 11 \end{array} \right\}. \qquad (1b)$$

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, $1 \le j \le 6$, 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, $6 < j \le 11$, 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 $\dot{\psi}_t$ 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 voiced phonation.

The fundamental amplitude A_t is assumed to be related to loudness perception [11] 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 $\overline{A}_{i,t}$ of a synthesis suited set of subbands,

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

The weights g_j are proportional to the 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 $\nu = 0.3$ [12] and $\mu = 1$ [3, 11]. The synthesis suited set of subbands is generated by replacing the over-complete subband set $6 < j \le 11$ by a set $6 < j \le N$, which is spaced equidistantly on the logarithmic frequency scale with 4 filters per octave,

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

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 [13]. The upper panel of figure 1 shows the analysis window for a section 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 Φ_j are given in a partially unwrapped form, depending

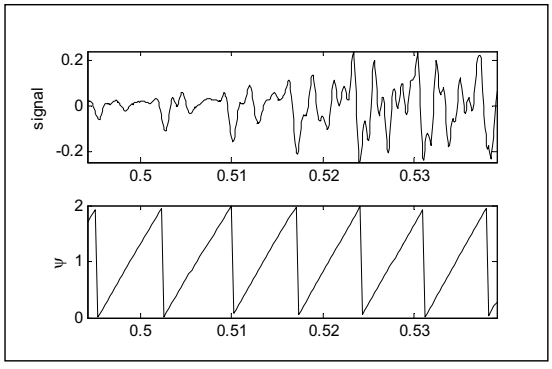

Figure 1, upper panel: 45 ms of a speech signal, which was taken from the /w/ in the word "wind" representing part of a publicly accessible pitch analysis data base [13]. The lower panel shows the reconstruction of the fundamental phase ψ in units of π. The time scale (in units of seconds) corresponds to the original one.

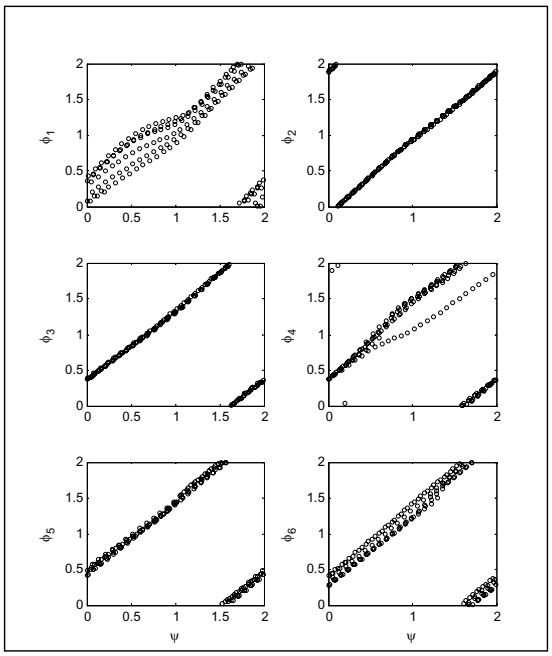

Figure 2: The relation of the subband phases Φ_j, $(j = 1,2,...,6)$, obtained from the speech signal of figure 1, to the fundamental phase ψ. The subbands 2, 3 and 5 are characterized by near perfectly linear phase relations, whereas the other subbands are found to be unsuited for the reconstruction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

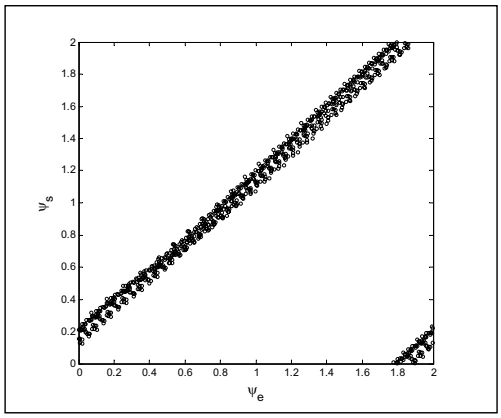

Figure 3: Relation between the wrapped up fundamental phase ψ_s, obtained from the speech signal, and the fundamental phase ψ_e, obtained from the electro-glottogram. Both fundamental phases are gained from 160 ms of uninterrupted voiced speech.

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 phases 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.

3. Entrainment of the primary response

In spite of the arbitrariness of the initial fundamental phase, the reconstruction of the glottal master oscillator is well suited to serve as fundamental drive of a two level drive – response model, which

can be seen as a natural extension of the well known source - filter model. The additional subsystem describes the vocal tract excitation as primary response of the fundamental drive [5-7]. The classical secondary response subsystem describes the signal forming, which results from resonances of the vocal tract. The subband decomposition (1) and (3), being used for the reconstruction of the fundamental drive, can also be used advantageously to achieve a numerically robust reconstruction of the excitation.

Due to the slow velocity of the glottal tissue (compared to the velocity of sound) the excitation $E_{j,t}$ of a voiced subband with index $1 \le j \le N$ can be assumed to be restricted (enslaved or entrained) to a generalized synchronization manifold (surface) in the combined state space of drive and response. In the simplest case the time dependence of subband excitation $E_{j,t}$ can be replaced by a dependence on the simultaneous state of the fundamental drive [14]. More generally, the dependence of the state of the primary response on the state of the fundamental drive may degenerate to a multivalued mapping, which can, however, be expressed as a unique function of the unwrapped fundamental phase ψ_t,

$$E_{j,t} \;=\; A_t\, G_{j,p}(\psi_t) \;=\; A_t \sum_{k \in S_{j,p}} c_{j,k}\, \exp(i\,k\,\frac{\psi_t}{p}). \qquad (4)$$

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_t and the potentially fast varying complex coupling function $G_{j,p}(\psi_t)$, which expresses the obvious richness of the excitation of human speech [5, 6]. In its general form, $G_{j,p}(\psi_t)$ represents a $2\pi\, p$ periodic function of the unwrapped fundamental phase ψ_t with an integer period number $p \ge 1$ and can thus be approximated by the finite Fourier series (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 [14], whereas $p = 2$ is suitable to describe the period doubling voice type [9]. 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 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 $\{X_{j,t}\}$ can thus be described by the following nonlinear conditional stochastic process with a two-level drive – response model as deterministic part (skeleton) [5-7],

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

where Δ denotes the subband specific prediction step length, b_j the complex resonator parameter, $\xi_{j,t}$ a (0,1) Gaussian complex white noise process and σ_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 $S_{j,p}$ of equation (4) is chosen in accordance to the respective bandpass filter. To avoid a bad conditioning of the parameter estimation in the case of a near periodic fundamental drive, the index set $S_{j,p}$ is pruned by the index, which equals the subband specific centre harmonic number h_j. The decomposition into subbands is used to estimate equation (5) with a subband specific time step length chosen as integer Δ, which approximates half of the ratio of the sample rate to the respective centre filter frequency F_j.

In the case of voiced speech segments, which contain sustainable voiced consonants, the continuous reconstructtuion of the fundamental phase can be used advantageously to extend equation (5) by a second excitation term $A_{t-\tau}\, G_{j,p}(\psi_{t-\tau})$ with a coupling function, which depends on a delayed fundamental phase. According to Teager and Teager [15] the delay τ 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. It cannot be excluded that the delay exceeds the length of the analysis window, a fact which affirmatively invalidates the assumption of stationary excitation. 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.

4. Properties of the excitation used as cue to distinguish voiced syllables

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. With the help of three examples it is demonstrated, how properties of the aggregated coupling functions can be used to infer physical details of the excitation process.

As a characteristic feature of vowels the time point of the glottal closure can be detected as a unique pulse (or as a unique outstanding slope) (figure 5). 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. This anchoring of the fundamental phase with the help of the vowels sheds new light on the old question, whether phonemes or syllables should be considered as the atoms of speech.

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 change of the phase position of the glottal pulse [16]. Figures 4 and 5 demonstrate the transition from the vowel to the nasal in the word “wind”, which has been used in figure 3. The two figures reveal a phase shift of about 1/12 of the fundamental cycle. Since figure 3 fully covers the respective time interval, it can be excluded that the observed phase shift results from an erroneous reconstruction of the fundamental phase. Together with the known 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 [16] 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.

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 6). 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

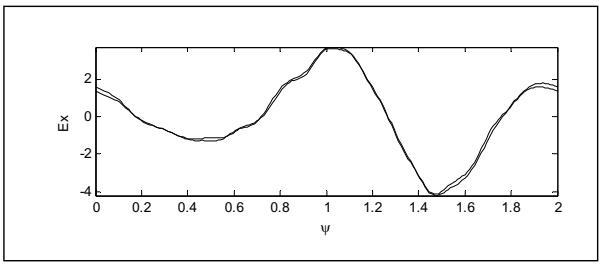

Figure 4: Aggregated fundamental phase dependent coupling function, reconstructed with period $p = 2$ for the vowel of the first occurrence of the word "wind" uttered by the first female speaker. The two curves correspond to the odd and even periods. The good agreement can be interpreted as a hint to the high robustness of the deconvolution.

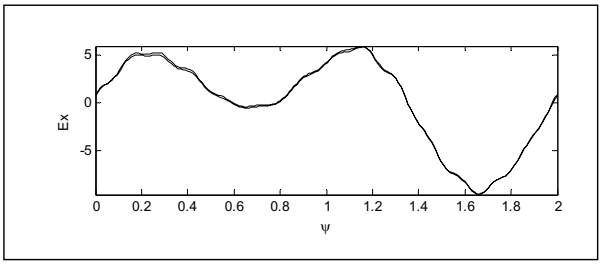

Figure 5: Aggregated fundamental phase dependent coupling function, reconstructed with period $p = 2$ for the nasal of the word "wind" of figure 4.

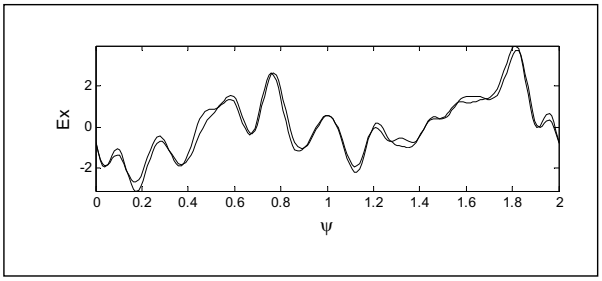

Figure 6: Aggregated fundamental phase dependent coupling function, reconstructed with period $p = 2$ for the voiced approximant /l/ of the word "along" uttered by the first male speaker. The disagreement between the odd and even curve shows a marked dependence on the fundamental phase.

dependence on the speaker and on the fundamental phase. 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 seconddary constrictiton of the vocal tract. The coupling functions with periodicity $p = 2$ are suited to describe a "voice type", which may be difficult to classifiy by using cycle lengths differences (figure 6). 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.

5. Discussion and conclusion

The present study is based on the link between speech-acoustics 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 turbulent airflow of voiced speech can be described by a low dimensional synchronisation manifold, is a remarkable hypothesis, which should be interpreted as result of the ontogenetic adaptation of the speech production. The success of the proposed description of non-pathological voiced human 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 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.

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

6. 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] Kawahara H., I. Masuda-Katsuse and A. de Cheveigné, *Speech Communication* **27**, 187-207 (1999)
[6] Drepper F.R. in C. Manfredi (editor), *MAVEBA 2003*, Firenze University Press (2004)
[7] Drepper F.R., *Fortschritte der Akustik-DAGA'05*, a and b (2005)
[8] Drepper F.R., *Phys. Rev. E* **62**, 6376-6382 (2000)
[9] Titze I.R., *Acta Acustica* **90**, 641-648 (2004)
[10] Hohmann V., *Acta Acustica* **10**, 433-442 (2002)
[11] Moore B.C.J., *An introduction to the psychology of hearing*, Academic Press (1989)
[12] Sottek R., *Modelle zur Signalverarbeitung im menschlichen Gehör*, Verlag M. Wehle, Witterschlick/Bonn (1993)
[13] ftp.cs.keele.ac.uk/pub/pitch
[14] Rulkov N.F. , M.M. Sushchik, L.S. Tsimring, H.D.I. Abarbanel, *Phys. Rev. E* **51**, 980-994 (1995)
[15] Teager H. M. and S. M. Teager, "Evidence for nonlinear sound production in the vocal tract," in *Proc NATO ASI on Speech Production and Speech Modelling*, pp. 241–261 (1990)
[16] Kawahara H. and P. Zolfaghari, *Eurospeech 2001* (2001)