# Selfconsistent time scale separation of instationary speech signals

**F.R. Drepper**

Forschungszentrum Jülich GmbH, 52349 Jülich, f.drepper@fz-juelich.de

The vocal tract excitation of a voiced speech signal results from a pulsatile airflow, which is strongly coupled to the dynamics of the vocal fold [1]. Due to the pronounced mass density differrence the coupling between the airflow and the glottal tissue is characterized by a dominant direction of interaction, such that the glottal oscillator can be assumed to take the role of the drive of the vocal tract excitation.

The frequency fluctuations of the glottal oscillator show a band limitation at the high frequency end and are known as an ubiquitious and important ingredient of speech communication. Whereas speech coding treats the acoustic excitation as an instationary process, speech analysis so far assumes the voiced excitation more or less explicitly to be stationary during the extent of the analysis window, usually chosen as 20 ms [2, 3].

The present analysis treats the broadband excitation as a stationary response of an instationary, narrowband fundamental drive, which is extracted selfconsistently from voiced sections of the speech signal. The selfconsistency includes the confirmation of its role as a topologically equivalent reconstruction of a glottal master oscillator, which synchronizes the acoustic excitation [4, 5].

To analyse synchronization or mode locking phenomena between instationary subsystems it is useful to determine the phases of bandlimited oscillators [6]. In contrast to the well known speech analysis, which considers time independent phases of Fourier components with a zero bandwidth, the present approach is focussed on time dependent phases of subbands with finite bandwidths. To determine the latter type of phases it is useful to describe all subbands and oscillators by complex variables.

### Extraction of the fundamental drive

As an important property of voiced human speech the response related state of the fundamental drive is assumed to be described uniquely by a fundamental amplitude, which is related to loudness perception and a fundamental phase, which is related to pitch perception. The amplitude and phase of the fundamental drive are extracted from a subband decomposition of the speech signal based on 4th order complex gammatone bandpass filters with approximately audiological bandwidths and with a constant analysis - synthesis delay as described in Hohmann [7].

The extraction of the fundamental phase $\psi_t$ relies on centre filter frequentcies $F_j$ of the subband decomposition, which are iteratively adapted to the frequency of the glottal master oscillator and its higher harmonics.

- In deviation from Hohmann [7] the audiological ERB scale has been replaced by a piecewise linear and logarithmic equivalent rectangular bandwidth (erb) scale, which is equidistantly spaced on the linear frequency scale at the lower frequency end and logarithmically spaced at the higher frequency end,

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

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

- As a second feature of human speech it is assumed that voiced phones are produced with at least two subbands in the lower harmonic (separable) range, which are not distorted by vocal tract resonances or additional constrictions of the airflow. For sufficiently adapted centre filter frequencies these subbands show a near linear (n:m) phase locking. The corresponding linear phase relations can be interpreted to result from near linear (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.
- 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 $\psi_t$ with a high time and frequency resolution.

The fundamental amplitude $A_t$ is assumed to be related to the loudness perception [8] by a power law. The exponent $\nu$ is chosen in such a way that the fundamental amplitude represents a linear homogenous function of the averaged subband amplitudes $\overline{A}_{i,t}$,

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

The weights $g_j$ are proportional to inverse hearing thresholds. In the range up to 3 kHz they can roughly be 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 $\nu = 0.3$ [9], $\mu = 1$ [3, 8] and an unconventional normalisation of the loudness!

The introduction of the fundamental drive leads to a useful time scale separation, which separates the high frequency phenomena of speech signals in the frequency range above the pitch from the ones below the pitch. Whereas a companion study [10] describes the estimation of a stationary model of the high frequency phenomena, the present study is focussed on that part of the low frequency phenomena, which is relevant for the instationarity of the high frequency phenomena. The main problem of conventional speech analysis lies in the fact that the amplitude and phase velocity of the instationary glottal oscillator cannot be assumed to be constant during a time window, which is long enough to get a good estimate of the stationary model of the respective high frequency process. The present study demonstrates that human speech offers a solution out of this dilemma.

### Examples from a pitch analysis data base

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 (very useful) pitch analysis database of the Keele University [11]. The upper trace of figure 1 opens an analysis window of 45 ms for a speech segment taken from the /w/ in wind spoken by the first male speaker as part of the sentence "The North wind and the sun …". The lower trace shows the reconstruction of the fundamental phase (given in wrapped up form), based on the subbands with the harmonic numbers 2, 3 and 5. The near perfectly linear phase synchronization of these subbands, which is used for the reconstruction of the drive, is demonstrated in figure 3, which shows the correspondding phase relations to the fundamental phase. The subband phases $\Phi_j$ 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. Figure 2 shows the analogues of figure 1 obtained from the simultaneous recording of the electro-glottogram. The comparison between the figures 1 and 2 underlines the exchangeability of the two fundamental phases. Note that even six linearly related subband phases are not suited to determine a unique initial phase of the glottal oscillator.

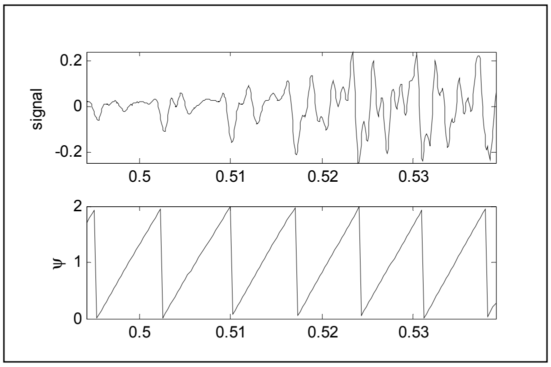

**Fig. 1**, upper trace: 45 ms of a speech signal, which was taken from the /w/ in the word "wind" representing part of a publicly accessible pitch analysis data base [11]. The lower trace shows a reconstruction of the fundamental phase $\psi$, obtained from the subbands 2, 3 and 5. Phases are given in units of $\pi$. The time scale corresponds to the original one and is given in units of seconds.

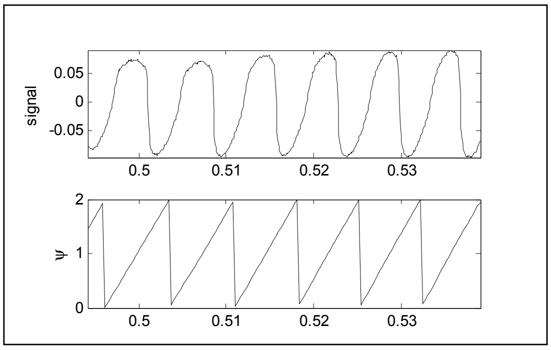

**Fig. 2.**, upper trace: the electro-glottogram, which was recorded simultaneously with the speech signal of figure 1. The corruption of the zero level has been reduced by applying a 75 Hz (moving average) high pass filter. The lower trace shows the reconstruction of the fundamental phase $\psi$, obtained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

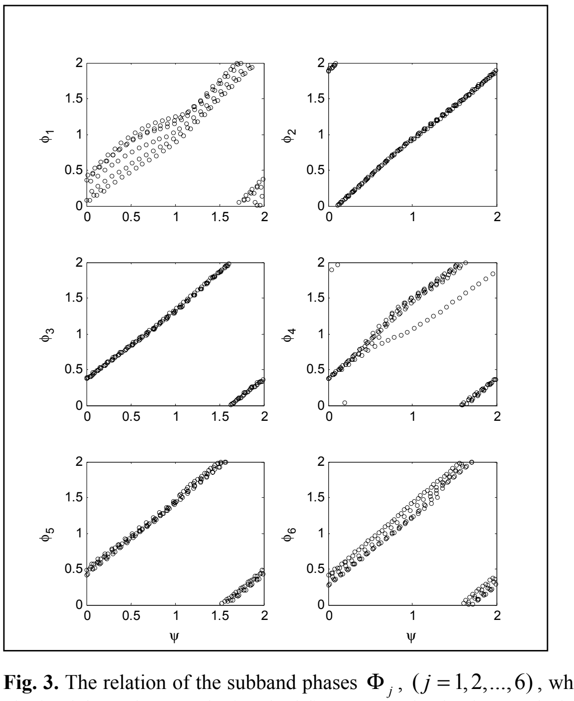

**Fig. 3.** The relation of the subband phases $\Phi_j$, $(j = 1, 2, ..., 6)$, which are obtained from the speech signal of figure 1, to the fundamental phase $\psi$. The subbands 2, 3 and 5 are characterized by near perfectly linear phase relations, whereas the other subbands turn out to be unsuited for the reconstruction of the fundamental phase.

A second demonstration of the exchangeability of the two drives is given in figures 4 and 5, which show the result of the pitch extraction for 100 successive analysis windows covering the section "North wind and the sun". In each figure the upper trace indicates the number of phase synchronous subbands, which could be used for the reconstruction of the fundamental phase and the lower trace indicates the adapted centre filter frequency of the fundamental subband, which results from the iterative adaption. In the case of voiced sections the filter frequency of the first subband coincides with the average phase velocity of the fundamental drive.

In the companion study it is shown that the reconstruction of the master oscillator of voiced excitation can successfully be used to reconstruct the voiced excitation and that the reconstructed voiced excitation can be used to resolve the arbitrariness of the initial fundamental phase [10].

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, London, P. Grassberger, H. Halling, M. Schiek and P. Tass, Jülich for helpful discussions.

**References**:

[1] Fant G. *Acoustic theory of speech production*, Mouton, 'S-Gravenhage (1960)
[3] Schroeder M.R., *Computer Speech*, Springer (1999)
[4] Drepper F.R., *Fortschritte der Akustik-DAGA'03*, (2003)
[5] Drepper F.R. in C. Manfredi (editor), *MAVEBA 2003*, Firenze University Press (2004)
[6] Drepper F.R., *Phys.Rev.E* **62**, 6376-6382 (2000)
[7] Hohmann V., *Acta Acustica* **10**, 433-442 (2002)
[8] Zwicker E. und Feldtkeller R., *Das Ohr als Nachrichtenempfänger*, Hirzel Verlag, (1967)
[9] Sottek R., *Modelle zur Signalverarbeitung im menschlichen Gehör*, Verlag M. Wehle, Witterschlick/Bonn (1993)
[10] Drepper F.R., *Fortschritte der Akustik-DAGA'05*, (2005)
[11] ftp.cs.keele.ac.uk/pub/pitch

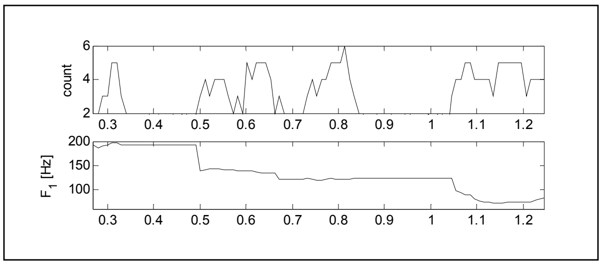

**Fig. 4.** Results of the pitch tracking for 100 analysis windows as shown in figure 1. The upper trace indicates the count of the number of near perfectly linear phase relations, which have been uncovered by the precise adjustment of the centre filter frequencies. The lower trace indicates the adjusted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

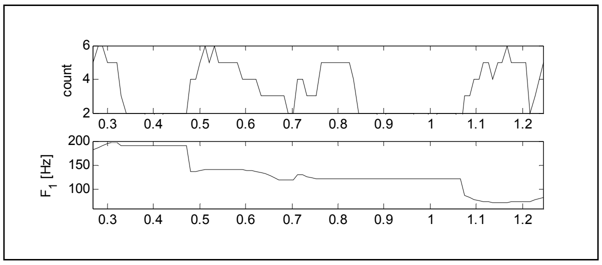

**Fig. 5.** Results of the pitch tracking for 100 analysis windows as described in figure 2. The comparison with figure 4 shows that the count of the number of perfectly phase synchronous subbands is generally higher in figure 5 in spite of the fact, that the electro-glottogram is corrupted by fluctuations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