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

F.R. Drepper

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

Vocal tract excitation of a voiced speech signal results from a pulsatile airflow, which is strongly coupled to the dynamics of the vocal fold. The excitation is generated either directly in the vicinity of the vocal fold and/or as acoustic (high frequency) part of a pulsatile, intermittently turbulent airflow in the vicinity of secondary constrictions of the vocal tract [1]. Due to 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 be the driving subsystem.

Though the frequency fluctuations of the glottal oscillator are known as an ubiquitious and important ingredient of speech communication, speech analysis so far assumes the voiced excitation more or less explicitly to be (weakly or wide sense) stationary during the extent of the analysis window, usually chosen not larger than 20 ms [2, 3].

The present analysis treats the broadband excitation as a stationary response of an instationary, bandlimited fundamental drive, which is extracted selfconsistently from voiced sections of the speech signal. A companion study [4] demonstrates that at least a useful first approximation of the fundamental drive can be extracted directly from a voiced speech signal (without the need of a simultaneous reconstruction of the excitation). The extraction relies on a voice specific subband decomposition and exploits the fact that the articulatory signal forming leaves a subset of the subband phases undistorted.

The present study is focussed on the analysis of the synchronization or mode locking phenomena, which result from the obviously nonlinear coupling of the vocal tract excitation to the band limited glottal master oscillator. Since the excitation cannot be observed directly, the analysis has to be based on a two-level drive - response model with a nonlinear primary response and a linear secondary response, known from the source and filter model.

Entrainment of the primary response

Due to the slow velocity of the glottal tissue (compared to the velocity of sound) the subband excitation $E_{j,t}$ of a voiced subband with an index $1 \leq j \leq N$ can be assumed to be restricted (enslaved or entrained) to a centre manifold in the combined state space of drive and response, in particular to a synchronization manifold (coupling function), which represents the (primary) response as a unique function of the simultaneous state of the respective drive [5],

$$E_{j,t} \;=\; A_t\, G_{j,p}(\psi_t) \;=\; A_t \sum_{k \in S_{j,p}} c_{j,k}\, \exp(i\,k\,\frac{\psi_t}{p}). \qquad (1)$$

As part of an improved time scale separation [4] the centre manifold is assumed as 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 [6, 7]. In its most general form $G_{j,p}(\psi_t)$ represents a $2\pi\, p$ periodic function of the unwrapped fundamental phase ψ_t with the integer period number $p \geq 1$. 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 Rulkov case [5], whereas $p = 2$ is particularly suited for the analysis of small amplitude (micro) tremor. When $2\pi\, p$ exeeds the length of the analysis window, the fundamental phase becomes a good substitute for the time with the effect that equation (1) is suited to describe a fully general (unvoiced) excitation.

The excitation $E_{j,t}$ is potentially modified by articulatory signal forming in particular by resonances of the vocal tract, which are described as secondary linear response. The subband decomposition can be used with advantage to improve the robustness of the parameter estimation of the resulting two-level drive – response model. As a first step toward this aim, the secondary response is simplified by the assumption that each subband is influenced at most by one dominant (isolated) first order pole (Helmholtz) resonator. Under this assumption, we arrive at the following nonlinear conditional stochastic process with a two-level drive – response model as deterministic skeleton [6, 7],

$$X_{j,t+\Delta} \;=\; b_j\, X_{j,t} \;+\; A_t\, G_{j,p}(\psi_t) \;+\; A_t\, \sigma_j\, \xi_{j,t}\,, \qquad (4)$$

where $X_{j,t}$ denotes the (reconstruction of the) complex output of the filter bank, Δ the subband specific prediction step length, b_j the subband specific 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 Fourier coefficients $c_{j,k}$ and the resonator parameter b_j can be achieved by multiple linear regression. The summation index set $S_{j,p}$ of equation (3) is chosen in accordance to the respective bandpass filter. To avoid a bad condition number of the parameter estimation, the index set $S_{j,p}$ is pruned by the index, which approximates the subband specific centre harmonic number h_j [4]. The decomposition into subbands can also be used to reduce the corrupting influence of the observational noise by estimating equation (4) with integer, subband specific step length Δ, chosen as approximation of half of the ratio of the sample rate to the respective centre filter frequency. Together with the extension of the analysis window (to 35 ms) these precautions lead to an unprecedented robustness of the reconstruction of the voiced excitation.

Properties of the two-level drive – response model as cues to distinguish voiced phonemes or 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 voiced consonants in particular of sustainable ones. The number of undisturbed subbands and/or the number of subbands, which can be described well by a low period coupling function is often found to be higher in the case of voiced consonants. The centre harmonic number of the latter type subbands reaches sometimes higher than 20. 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 and filter model. When the described nonlinear deconvolution is applied to voiced sections of a speech signal it turns out that the aggregated coupling function shows a large amount of detail, which can be reconstructed with a surprising time invariance in the case of sustainable voiced phones. With the help of three examples it is demonstrated, that details of the coupling function can be associated with phoneme specific excitation events.

In the case of the voiced approximant /l/ the aggregated coupling function shows several steep slopes, which act as an aperiodicity amplifier on the week but important jitter of the fundamental drive (figure 1). The successful deterministic description of the aperiodicity of the sustainable voiced consonants demonstrates that the turbulence at their phoneme specific constrictions has a causal link to the glottal dynamics, which generates a sensitive dependence on the phase of the glottal master oscillator. The turbulent brake up of the symmetry of the laminar flow below the glottis obviously depends in a sensitive way on details of the glottal fold and/or of the supra-glottal air duct. The pattern of the aggregated coupling function is thus expected to contain cues for phoneme and for speaker recognition. First results show that the extent of the deterministic aperiodicity amplification shows a marked dependence on the speaker and on the fundamental phase.

Vowels are characterized by the fact that the time point of the glottal closure event can be detected as a unique single pulse (or as a single outstanding slope) (figure 2). Since there is hardly any syllable without a vowel, these time points can be used to resolve the arbitrariness of the initial fundamental phase [4] and to calibrate the wrapped up fundamental phase in terms of the time interval since the last closure event. The fundamental phase dependence of the aggregated coupling function can thus be interpreted in terms of a run time spectroscopy. The anchoring of the fundamental phase with the help of the vowels sheds new light on the old question, whether syllables or phonemes should be considered as the atoms of speech.

Due to the different length of the nasal tract in comparison to the one of the vocal tract, nasals can be distinguished by a (sudden) change of the phase position of the glottal pulse (figure 3). A second cue is a characteristic second pulse, which represents the echo from the vocal tract and can be used to distinguish the /n/ from the /m/.

The distinction of different vowels is well known to rely on resonator properties of the vocal tract (Vary et al. 1998, Schroeder 1999), which can be determined with the help of parameter b_j of equation (4a). The improved robustness of the estimation of the set of parameters b_j can also be used to explore transient secondary responses with $|b_j| \neq 1$, which represent the second type of instationarity, which can be described with the help of the two-level drive – response model. Subband specific, transient secondary responses are expected to be useful cues for numerous syllables, which combine a voiced stop consonant and a vowel like in “bee” [8].

For strictly periodic voiced phones the estimation of the two-level drive – response model shows a degeneracy (divergence of the confidence intervals of the resonator parameters) due to a break down of the distinction between properties of the first level coupling and the ones of the second level. This finding sheds new light on the role of the ubiquitious frequency fluctuations of the glottal master oscillator.

Summary

The transmission protocol of voiced human speech is known to be based on the production and analysis of turbulent airflow pattern in the extended 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 voiced airflow by replacing the time dependent excitation of the classical source and filter model by a fundamental phase dependent coupling function, which can be interpreted as a one dimensional generalized synchronization manifold of a turbulent airflow pattern with a tamed complexity. To make the synchronization manifold visible (or audible), it is necessary to introduce a voice specific subband decomposition of the speech signal and to extract a fundamental drive from the speech signal. The evolution of speech has lead to many voiced phones and syllables, which can be distinguished by properties of this one dimensional synchronization manifold and the closely related two-level drive - response model. Furthermore the phase dependence of the coupling function shows marked interindividual differences. Nonpathological voiced speech can be expected to support the reconstruction of the fundamental drive by leaving at least two subbands of the excitation undistorted by vocal tract resonance and/or secondary constriction.

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] Drepper F.R., *Fortschritte der Akustik-DAGA’05*, (2005)
[5] Rulkov N.F. , M.M. Sushchik, L.S. Tsimring, H.D.I. Abarbanel, *Phys. Rev. E* **51**, 980-994 (1995)
[6] Drepper F.R., *Fortschritte der Akustik-DAGA’03*, (2003)
[7] Drepper F.R. in C. Manfredi (editor), *MAVEBA 2003*, Firenze University Press (2004)
[8] Moore B.C.J., *An introduction to the Psychology of hearing*, Academic Press (1989)

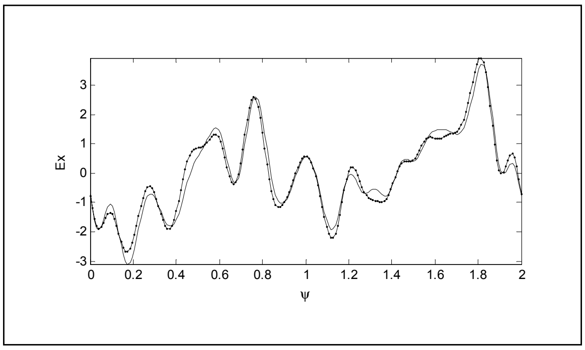

Fig. 1: aggregated fundamental phase dependent coupling function (synchronization manifold) reconstructed with period $p = 2$ for the voiced approximant /l/ of the word “along” uttered by the first male speaker as part of the pitch analysis data base of the Keele University. The two curves correspond to the odd and even periods. The good agreement can be interpreted as a hint to the high robustness of the deconvolution.

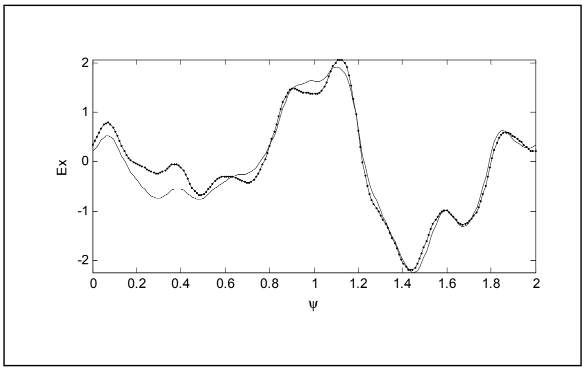

Fig. 2: aggregated fundamental phase dependent coupling function reconstructed with period $p = 2$ for the vowel of the first occurrence of the word “sun” uttered by the second male speaker.

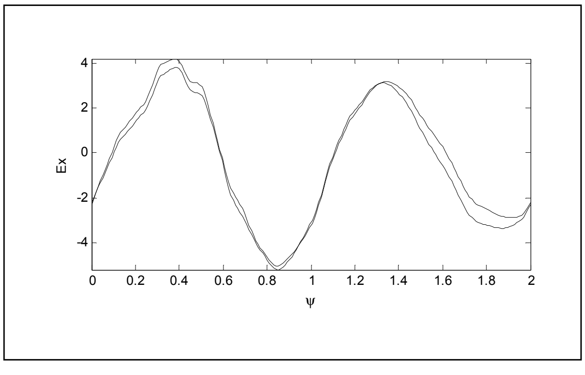

Fig. 3: aggregated fundamental phase dependent coupling function reconstructed with period $p = 2$ for the nasal of the word “sun” of figure 2.