

Non-stationary, filter-stable acoustic objects as atoms of voiced speech

Friedhelm R. Drepper¹, Ralf Schlüter²

¹ *Zentralinstitut für Elektronik, Forschungszentrum Jülich, Email: f.drepper@fz-juelich.de*

² *Lehrstuhl für Informatik 6, Computer Science Department, RWTH Aachen University*

In spite of the undisputedly high degree of non-stationarity of speech signals, the present day determination of its acoustic features is based on the assumption that *speech production* can be described as a linear time invariant (LTI) system on the time scale of about 20 ms [1]. In automatic speech recognition, the wide sense stationarity of an LTI-system is used as prerequisite for the consistent estimation of Fourier spectra or of autoregressive models [1]. As an evolutionarily plausible supplement, *speech perception* is also assumed to be focussed on acoustic features which are obtained by using the LTI assumption [2]. Present day models of pitch perception are no exception [2, 3].

Speech signals are also characterized by a huge range of relevant time scales. (Virtual) pitch perception introduces a basic time scale separation which separates the phonological and linguistic time scales from the phonetically relevant ones. The present study is based on the hypothesis that pitch perception is able to achieve this separation without relying on a frequency gap in the long term spectrum (being equivalent to the assumption of temporary stationarity).

As has been pointed out by Terhardt [3], pitch perception can be trained to switch between analytic listening to spectral pitch and synthetic listening to virtual pitch. As an important result of the present study, the re-engineering of spectral pitch perception leads to the reconstruction of empirical modes which can be interpreted as topologically equivalent images of more or less resonantly excited oscillatory (formant) modes on the transmitter side. Furthermore it is shown that voiced phones support a fast converging iterative reconstruction of the formant modes on the receiver side and that the (filter) stable reconstruction of the formant modes is well suited for a phase modulation transmission protocol of voiced speech.

Non-stationary bandpass filters

Hearing models are well known to use a subband decomposition with so called critical (audiological) bandwidths [2, 4]. These bandwidths limit the number of separable harmonics to the range 6-8. If we describe

the input signal as $S(t)$, the envelope of the impulse response of the bandpass filter as $W_j(t)$, the distance of the envelope maximum to the start of the response as τ_j and the centre filter frequency (CFF) as $\omega_j/2\pi$, we obtain a complex subband of the form

$$X(\omega_j, t) = \int_{-\infty}^{t+\tau_j} d\tau S(\tau) W_j(t-\tau) e^{i\omega_j(t-\tau)}, \quad (1)$$

which represents an analytical signal (due to the limited bandwidth) and is thus well suited to obtain subband phases $\varphi_j(t)$. An autoregressive approximation to a gammatone bandpass filter represents a widely used example [4-7]. The CFFs $\{\omega_j | j=1, \dots, J\}$ are usually chosen with the aim to achieve approximate orthogonality and completeness of the decomposition [4]. In deviation to common hearing models we start the analysis with CFFs of virtual pitch perception models [2, 3], i.e. with CFFs which form a harmonic (equidistant) grid (template) for the harmonics 1-8. Deviating from common pitch models we introduce a time dependent CFF into *subband* (1) [6, 7]

$$X_j(t) = \int_{-\infty}^{t+\tau_j} d\tau S(\tau) W_j(t-\tau) \exp(i \int_{\tau}^t \omega_j(\tau') d\tau') \cdot \quad (2)$$

To demonstrate useful properties of *part-tone* (2), we choose an input signal which is often termed as *sinusoid*

$$S(\tau) = A(\tau) \cos \Phi_j(\tau) \quad \text{with} \quad (3a)$$

$$\Phi_j(\tau) = \int_0^{\tau} \Omega_j(\tau') d\tau', \quad (3b)$$

$$A(\tau) > 0 \quad \text{and} \quad \Omega_j(\tau) > 0, \quad (3c)$$

i.e. amplitude $A(\tau)$ and phase velocity $\Omega_j(\tau)$ are both limited to positive values. In analogy to phase definition (3b), we define filter phase $\psi_j(\tau)$ by a time integral of the respective CFF $\omega_j(\tau)/2\pi$. For times $t \gg \tau_j$, analytical signal (2) simplifies to

$$X_j(t) = \frac{1}{2} e^{i\psi_j(t)} \int_0^{t+\tau_j} d\tau A(\tau) W_j(t-\tau) e^{i(\Phi_j(\tau) - \psi_j(\tau))}.$$

Choosing the special case $\omega_j(\tau') = \Omega_j(\tau')$, i.e. a CFF contour which is precisely adapted to the sinusoid, and assuming a slowly varying amplitude $A(\tau)$, we get

$$X_j(t) = \frac{1}{2} A(t) e^{i\Phi_j(t)}, \quad (5)$$

i.e. a part-tone which represents the analytic signal of the input signal. Since equation (5) is valid for arbitrary times, result (5) can also be interpreted in a different way. If (in case of a slowly varying amplitude) we succeed to adapt the CFF of a part-tone to its instantaneous frequency, we are assured to have reconstructed a part-tone with an uncorrupted phase [6, 7].

For a given CFF contour all input signals with a different phase velocity experience a damping due to interference. A filterbank with bandpass filters of form (2) or (4) is thus well suited to separate several empirical modes with different frequencies without corrupting their phases. However, there remains the problem of finding the appropriate CFF contours. In this situation we hypothesize that voiced speech supports a robust and efficient adaptation algorithm on the receiver side and that the adaptation of the CFFs can be achieved iteratively by using the instantaneous frequency contours of the respective part-tones.

Stability of Centre Filter Frequencies

To analyze the convergence of the iterative adaptation of a CFF contour it is useful to approximate the envelope of the impulse response in equation (4) by a Gaussian

$$W_j(t-\tau) = W_j(0) \exp\left(-\frac{\alpha_j}{2}(t-\tau)^2\right) \quad (6)$$

where α_j denotes the inverse of its variance. If we assume amplitude $A(\tau')$ as slowly varying in comparison to envelope (6), part-tone (4) simplifies to

$$X_j(t) = W_j(0) A(t) e^{i\psi_j(t)} \int_{-\infty}^{\infty} d\tau' e^{-i(\psi_j(t+\tau') - \Phi_j(t+\tau'))} e^{-\frac{\alpha_j}{2}\tau'^2}.$$

Furthermore, we assume a smooth instantaneous phase velocity $\Omega_j(\tau')$ of the sinusoid (oscillatory subsystem on the transmitter side). Within a sufficiently short analysis window, the phase of the sinusoid can therefore be described well by a quadratic function. In case of the CFF, we deliberately choose a linear time trend within the current window. The phase difference in equation (7) can thus be written as

$$\begin{aligned} \psi_j(t+\tau') - \Phi_j(t+\tau') &= \psi_j(t) - \Phi_j(t) + (\omega_j(t) - \Omega_j(t)) \tau' \\ &\quad + (\dot{\omega}_j(t) - \dot{\Omega}_j(t)) \tau'^2 / 2 \\ &= \Delta\psi_j(t) + \Delta\omega_j(t) \tau' + \Delta\dot{\omega}_j \tau'^2 / 2, \end{aligned} \quad (8)$$

where we have introduced useful abbreviations in the last step. In case of the error term $\Delta\dot{\omega}_j$ of the CFF chirp, we can drop the time argument in view of the above assumptions. As a further benefit, the integrand in equation (7) simplifies to a complex Gaussian. When the integral is solved analytically, we arrive at phase $\varphi_j(t)$ of part-tone (7)

$$\varphi_j(t) = \Phi_j(t) + \frac{1}{2} \Delta\dot{\omega}_j \frac{\Delta\omega_j(t)^2}{\alpha_j^2 + \Delta\dot{\omega}_j^2} - \frac{1}{2} \arctan\left(\frac{\Delta\dot{\omega}_j}{\alpha_j}\right). \quad (9)$$

To make the convergence property of the iteration visible, we represent the phase error $\varphi_j(t) - \Phi_j(t)$ of the part-tone as function of the error of the respectively used CFF contour, and express this error as function of $\Delta\omega_j(t_0)$ and $\Delta\dot{\omega}_j$, where t_0 denotes the start of the current analysis window. Replacing $\Delta\omega_j(t) = \Delta\omega_j(t_0) + \Delta\dot{\omega}_j(t-t_0)$ in equation (9), we obtain in leading order of the small quantities $\Delta\omega_j(t_0)$ and $\Delta\dot{\omega}_j$

$$\begin{aligned} \varphi_j(t) - \Phi_j(t) &= \varphi_j(t_0) - \Phi_j(t_0) + \Delta\omega_j(t_0) \frac{\Delta\dot{\omega}_j^2}{\alpha_j^2} (t-t_0) \\ &\quad + \frac{\Delta\dot{\omega}_j^3}{\alpha_j^2} (t-t_0)^2 / 2 + \dots \end{aligned} \quad (10)$$

The left hand side can alternatively be expanded into a Taylor series for small $(t-t_0)$. When comparing the expansion coefficients with the corresponding terms of equation (10), we arrive at

$$\begin{aligned} \dot{\varphi}_j(t_0) - \dot{\Phi}_j(t_0) &= \Delta\omega_j(t_0) \Delta\dot{\omega}_j^2 / \alpha_j^2 \quad \text{and} \\ \ddot{\varphi}_j(t_0) - \ddot{\Phi}_j(t_0) &= \Delta\dot{\omega}_j^3 / \alpha_j^2. \end{aligned} \quad (11)$$

In its simplest form, the iteration replaces

$$\begin{aligned} \Delta\omega_j(t_0) &= \dot{\varphi}_j(t_0) - \dot{\Phi}_j(t_0) \quad \text{and} \\ \Delta\dot{\omega}_j &= \ddot{\varphi}_j(t_0) - \ddot{\Phi}_j(t_0). \end{aligned}$$

The described iteration converges with the *third power* of the deviation from a stable fixed point, i.e. faster than the well known Newton method and therefore faster than all optimization algorithms. As shown in equation (5), filter stability implies uncorrupted phases of the respective part-tone. For part-tones with several relevant harmonics, the sinusoid assumption being specified in equation (3c) is no longer valid. For such part-tones, convergence property (11) does not apply. Such part-tones are well known to have lower priority for virtual pitch perception [2,3] and are out of the scope of the present summary.

Re-engineering of virtual pitch perception

Voiced phones are known to have at least one formant in the range of the separable harmonics. In view of the additional existence of anti-formants, we must be prepared to find several stable fixed points of the iteration. Part-tones which converge to the same stable fixed point, define an empirical mode specific cluster. Within each cluster a second type of distance of a part-tone to the cluster centre can be defined as the number of iterations being necessary to converge. Part-tones with low distances to a cluster centre are synthesized [4] to an empirical mode which is consistent with the sinusoid assumptions (3c). The frequency of a formant mode is interpreted as acoustic correlate of spectral pitch [3].

For lower harmonic part-tones the vocal tract causes phase shifts, which introduce non-negligible frequency shifts in a non-stationary setting. This makes re-engineering of virtual pitch perception more difficult [2, 3]. As part of auditory scene analysis, a subset of empirical modes is selected, which is consistent with the current estimate of the fundamental frequency contour $f_0(t)$. As motivated in [5, 6], the selected empirical modes are interpreted as different transients of a two-level cascaded response system which is driven by a common (hidden) fundamental drive in the frequency range of the pitch. For the single harmonic dominated modes, the excitation(s) of the well known secondary (vocal tract) response are described as periodic functions of the *fundamental phase*. (For the other modes their envelopes are described as periodic functions.) The band-limited periodic functions can be represented by finite Fourier series. The empirical modes and the present estimate of $f_0(t)$ are used to estimate the Fourier coefficients and the resonator parameters simultaneously by multiple linear regression. For given excitation- and resonator parameters, the analysis window specific chirp rate $\dot{f}_0(t_0)$ of the fundamental drive is estimated by 1-dim. nonlinear regression. The chirp rate is used to update the fundamental frequency contour $f_0(t)$.

The stable $f_0(t)$ can be interpreted as acoustic correlate of virtual pitch and the resonator and excitation parameters can be used for the distinction of sustainable voiced phones. The description of the *primary* response as synchronized deterministic and/or synchronously modulated stochastic response with short autocorrelation leaves the *secondary* response as aggregate and/or approximation of all subsystems which generate *long* transients [5-7]. This alternative interpretation of the classical source-filter model is sufficiently general, to resolve the well known (but often ignored) contradiction of a time invariant vocal tract filter in case of active phonation. For a sufficiently broadband primary response, the cascaded response of a non-stationary fundamental drive generates non-stationary acoustic objects, which are suited to be analyzed by the outlined pitch- and vowel perception oriented empirical mode decomposition and time-scale separation.

References

- [1] Gold B. and N. Morgan, "Speech and audio signal processing", John Wiley & Sons", Chichester (2000)
- [2] Moore B.C.J., "An introduction to the psychology of hearing", Academic Press, London (1989)
- [3] Terhardt, E., Stoll, G., Seewann, M., "Algorithm for extraction of pitch and pitch salience from complex tonal signals", J. Acoust. Soc. Am. 71, 679-688 (1982)
- [4] Hohmann V., "Frequency analysis and synthesis using a gammatone filterbank", Acta Acustica 10, 433-442 (2002)
- [5] Drepper F.R., „Stimmhafte Anregung als synchronisierte Antwort eines eindeutigen fundamentalen Treibers“, Fortschritte der Akustik-DAGA'06 (2006)
- [6] Drepper F.R., „Voiced speech as response of a self-consistent fundamental drive“, Speech Comm. 49, 186-200 (2007)
- [7] Drepper F.R., "Non-stationary self-consistent acoustic objects as atoms of voiced speech", in M.Chetouani et al. (Eds), NOLISP 2007 LNAI 4885, 188-203, Springer, Heidelberg (2007)