Discrete evolutionary spectra and their application to a theory of pitch perception

Andreas Thumfart
Universität Heidelberg

July 29, 1995

Abstract

A definition of discrete evolutionary spectra is given that complements the notion of evolutionary spectral density given by Dahlhaus in [1]. For processes that have a discrete evolutionary spectrum, the asymptotic behavior of linear functionals of the periodogram is investigated. The results are applied in a mathematical analysis of Licklider's theory of pitch perception. A pitch estimator based on this theory is investigated with respect to the shift of the pitch of the residue described by Schouten et al. in [8].

1 Introduction

In his paper [1] Dahlhaus introduces a new notion of a *locally stationary process*. His approach differs from the well known one given by Priesley ([6, 7]) in being inherently asymptotic. This enables him to prove strong asymptotic results. Further, the spectral representation of a locally stationary process that Dahlhaus postulates is unique, in contrast to the one in Priestley's theory.

Dahlhaus defines a process to be locally stationary if it has a certain spectral representation. Every such process has an evolutionary spectral density. Hence his theory doesn't cover the case of a discrete spectrum. In section 2.1 of this paper we give a definition of a *locally stationary process with discrete evolutionary spectrum* and prove a uniqueness result for the spectral representation. In section 2.2 we discuss the asymptotic behavior of linear functionals of the periodogram of a process with discrete evolutionary spectrum. Finally we apply these results in section 3 to Licklider's theory of pitch perception (see [4]). We give a fast algorithm for a simplified version of his model and study its asymptotic behavior. A pitch estimator based on it is investigated with respect to the observations reported by Schouten et al. in [8].

2 The mathematical theory of discrete evolutionary spectra

2.1 Definition and some elementary properties

We define a process with discrete evolutionary spectrum as a process that can almost be written as a sum of pure oscillations. The amplitude, null-phase and frequency of every summand may change in time. But like in Dahlhaus' theory this change becomes slower and slower as the sample size increases. Here is the exact definition:

Definition 1 *A sequence of stochastic processes $X_{t,T}$ $(t = 1, \dots, T)$ is said to have a discrete evolutionary spectrum if*

1. *there exists a representation*

   $$X_{t,T} = \sum_{n \in M} A^0_{n,t,T} \, a.s.$$

   for some $M \subseteq \mathbf{Z}$ and

2. *for every $n \in M$ there exist a complex valued mean 0 stochastic process $A_n(u)$ on $[0,1]$ with a.s. differentiable paths and a sequence $\theta_{n,T}(t)$ $(t = 1, \dots, T)$ such that*

   $$\exists K \forall t, T \sum_n \left| A^0_{n,t,T} - A_n\left(\frac{t}{T}\right) \exp(i\theta_{n,T}(t)) \right| \leq KT^{-1} \ a.s.$$

   and

3. $$\exists K \forall u \sum_{n \in M} \sup_u |A_n(u)| \leq K \ a.s. \ ,$$
   $$\exists K \forall u \sum_{n \in M} \sup_u |A'_n(u)| \leq K \ a.s. \ ,$$

4. $$\forall u \in [0,1], n \in M : EA_n(u) = 0$$
   $$\forall u_1, u_2 \in [0,1], n \neq m, n, m \in M : Cov(A_n(u_1), A_m(u_2)) = 0$$

   and

5. *for every $n \in M$ there exists a function $\eta_n : [0,1] \to \mathbf{R}$ such that*

   $$\exists K \forall t, T, n \in M \ \left| \theta_{n,T}(t) - \theta_{n,T}(t-1) - \eta_n\left(\frac{t}{T}\right) \right| \leq KT^{-1}.$$

   η_n has a uniformly (in $n \in M$) bounded derivative $\frac{\partial \eta_n(u)}{\partial u}$.

$\eta_n(u)$ *is called an* instantaneous frequency *of* $X_{t,T}$ *at time* u*. We say that* $X_{t,T}$ *has a* spectral line *of hight* $Var(A_n(u))$ *at* $\eta_n(u)$*. When we deal with real valued processes, we always assume that* $\forall n \in M:\ -n \in M, 0 \notin M$ *and* $\forall u \in [0,1]\ A_n(u) = \overline{A_{-n}(u)}$ *and* $\forall t \leq T \theta_{n,T}(t) = -\theta_{-n,T}(t)$*.*

Example:

$$X_{t,T} = \sum_{n \in M} A_n\left(\frac{t}{T}\right) \exp\left(i\,\nu_n\left(\frac{t}{T}\right)t\right),$$

i.e., $\theta_{n,T}(t) = \nu_n(t/T)t$, where ν_n is twice continuous differentiable with bounded second derivative. Then we have $\eta_n(u) = \nu_n(u) - u\nu_n'(u)$. Note that in general we cannot choose $\eta_n(u)$ to be $\nu_n(u)$.

Proposition 1 *If* $X_{t,T}$ *is a process with discrete evolutionary spectrum as above then*

$$Cov(X_{t+\tau,T}, X_{t,T}) = \sum_n Var\left(A_n\left(\frac{t}{T}\right)\right)\exp\left(i\tau\eta_n\left(\frac{t}{T}\right)\right) + R,$$

where $|R| \leq O\left(\frac{1+\tau+\tau^2}{T}\right)$*.*

Proof By 2 and 3 of definiton 1 we may freely exchange expectations and the sum in expressions of the form $E\sum_{n\in M} A \ldots$. Therefore $X_{t,T}$ has a mean of order $O(1/T)$.

$$\begin{aligned}
E\overline{X_{t,T}}X_{t+\tau,T} &= E\sum_{n,m\in M}\overline{A^0_{n,t,T}}A^0_{m,t+\tau,T} \\
&= \sum_{n,m\in M} E\overline{A^0_{n,t,T}}A_m\left(\frac{t+\tau}{T}\right)\exp(i\theta_{m,T}(t+\tau)) + R_1 \\
&= \sum_{n,m\in M} E\overline{A_n\left(\frac{t}{T}\right)}A_m\left(\frac{t+\tau}{T}\right)\exp\left(i\left(\theta_{m,T}(t+\tau) - \theta_{n,T}(t)\right)\right) \\
&\quad +R_1 + R_2,
\end{aligned}$$

where $|R_1|$ and $|R_2|$ are of order $\leq O(1/T)$ a.s. by 2 of definition 1. Since A_n and A_m are uncorrelated for $n \neq m$ we get:

$$\begin{aligned}
&\sum_{n\in M} E\overline{A_n\left(\frac{t}{T}\right)}A_n\left(\frac{t+\tau}{T}\right)\exp\left(i\left(\theta_{n,T}(t+\tau) - \theta_{n,T}(t)\right)\right) \\
&+R_1 + R_2 \\
&= \sum_{n\in M} E\left|A_n\left(\frac{t}{T}\right)\right|^2 \exp\left(i\left(\theta_{n,T}(t+\tau) - \theta_{n,T}(t)\right)\right) \\
&+R_1 + R_2 + R_3,
\end{aligned}$$

$$\begin{aligned} R_3 &:= \sum_{n\in M} E\left(\overline{A_n\left(\frac{t}{T}\right)}A_n\left(\frac{t+\tau}{T}\right) - \left|A_n\left(\frac{t}{T}\right)\right|^2\right) \\ &\quad \exp\left(i\left(\theta_{n,T}(t+\tau) - \theta_{n,T}(t)\right)\right) \\ |R_3| &\leq O\left(\frac{\tau}{T}\right)\sum_{n\in M} Esup_u\left(|A_n'(u)||A_n(u)|\right) \end{aligned}$$

by the mean value theorem.

The next step illustrates a technique that is central to the theory of discrete evolutionary spectra. To obtain the result we replace $\theta_{n,T}(t+\tau)-\theta_{n,T}(t)$ by $\tau\eta_n(t/T)$. The error we get is

$$\begin{aligned} R_4 &:= \sum_{n\in M} E\left|A_n\left(\frac{t}{T}\right)\right|^2 \exp\left(i\tau\eta_n\left(\frac{t}{T}\right)\right) \\ &\quad \left(\exp\left(i\sum_{k=0}^{\tau-1}\left(\theta_{n,T}(t+\tau-k) - \theta_{n,T}(t+\tau-k-1) - \eta_n(t/T)\right)\right) - 1\right) \end{aligned}$$

Since

$$\left|\sum_{k=0}^{\tau-1}\left(\theta_{n,T}(t+\tau-k) - \theta_{n,T}(t+\tau-k-1) - \eta_n(t/T)\right)\right|$$

is $\leq O(\tau/T) + O(\tau^2/T)$ by 5 of definition 1 and the mean value theorem we have

$$|R_4| \leq O\left(\frac{\tau+\tau^2}{T}\right)\sum_{n\in M} E\left|A_n\left(\frac{t}{T}\right)\right|^2.$$

■

Therefore we define

Definition 2

$$F(u,\lambda) := \sum_{n\in M} Var(A_n(u))\,\mathbf{1}_{[\eta_n(u),\pi]}(\lambda)$$

is called spectral distribution function *of* $X_{t,T}$.

The spectral distribution function of a process with discrete evolutionary spectrum is uniquely determined by the covariance structure of the process:

Proposition 2 *Under the assumptions of proposition 1:*

$$\begin{aligned} &\lim_{K\to\infty}\frac{1}{2K+1}\sum_{\tau=-K}^{K}\lim_{T\to\infty} Cov\left(X_{[uT]+\tau,T}, X_{[uT],T}\right)\exp(-i\lambda\tau) \\ &= \sum_{n\in M} Var(A_n(u))\,\mathbf{1}_{\{\eta_n(u)\}}(\lambda) \end{aligned}$$

(The convergence is not uniform in λ*.)*

Here $[x]$ denotes the greatest integer $\leq x$.
Proof

$$\frac{1}{2K+1}\sum_{\tau=-K}^{K}\lim_{T\to\infty}\mathrm{Cov}\left(X_{[uT]+\tau,T},X_{[uT],T}\right)\exp(-i\lambda\tau)$$
$$=\sum_{n\in M}\mathrm{Var}\left(A_n\left(u\right)\right)\frac{1}{2K+1}\sum_{\tau=-K}^{K}\exp(i(\eta_n(u)-\lambda)\tau)$$

by proposition 1. The last sum is the dirichlet kernel, which is an approximate identity. ■

2.2 Linear functionals of the periodogram

Let $h:\mathbf{R}\to\mathbf{R}$ be a data taper and

$$d_N(u,\lambda):=\sum_{s=0}^{N-1}h\left(\frac{s}{N}\right)X_{[uT]-N/2+s+1,T}\exp(-i\lambda s)$$

the tapered fourrier transform of a segment of length N around $[uT]$ of the time series. N is assumed to be even.

$$H_{k,N}(\lambda):=\sum_{s=0}^{N-1}h\left(\frac{s}{N}\right)^k\exp(-i\lambda s),$$
$$I_N(u,\lambda):=\frac{1}{2\pi H_{2,N}(0)}d_N(u,\lambda)d_N(u,-\lambda).$$

We investigate the asymptotic behavior of functionals of the form

$$B_N(u,\phi):=\int_{-\pi}^{\pi}I_N(u,\lambda)\phi(\lambda)d\lambda,$$

where ϕ is a continuous 2π-periodic function. Let

$$B(u,\phi)\ :=\ \sum_{n\in M}|A_n(u)|^2\phi(\eta_n(u)).$$

Note that in general $B(u,\phi)$ is still random and $\neq\int_{-\pi}^{\pi}\phi(\lambda)dF(u,\lambda)$. If only the phase of $A_n(u)$ is random and the absolute value is deterministic, then $B(u,\phi)=\int_{-\pi}^{\pi}\phi(\lambda)dF(u,\lambda)$.
Assumption A1:

1. $X_{t,T}$ is a process with discrete evolutionary spectrum and

$$\forall u\in[0,1]\forall n\neq m\in M\ \eta_n(u)\neq\eta_m(u)\tag{1}$$

2. ϕ is a bounded, complex valued, continuous, 2π-periodic function.

3. The data taper $h : \mathbf{R} \to \mathbf{R}$ is of bounded variation.

4. For the segment length N and the sample size T, $(N^2 \log N)/T \to 0$ and $T/N^4 \to 0$ hold as $T \to \infty$.

Theorem 1 *Under assumption A1 the following holds: If a.s. there exists a $K < \infty$ such that for all $u \in [0,1]$*

$$\sum_{n \neq m \in M} \frac{|A_n(u)|\,|A_m(u)|}{|\eta_n(u) - \eta_m(u)|} \leq K \tag{2}$$

then

$$B_N(u,\phi) \to B(u,\phi) a.s. \ .$$

If there exists a $K < \infty$ such that for all $u \in [0,1]$

$$\sum_{n \neq m, l \neq k \in M} \frac{|E(A_n(u)\overline{A_m(u)A_l(u)}A_k(u))|}{|\eta_m(u) - \eta_n(u)|} \leq K \tag{3}$$

then

$$B_N(u,\phi) \to B(u,\phi)$$

in quadratic mean and in probability. The convergence is uniform in u in both cases.

The rest of this section contains the proof of theorem 1 and some technical tools that are needed for it. Let

$$H_N(f(\cdot),\lambda) := \sum_{s=0}^{N-1} f(s) \exp(-i\lambda s),$$

$$L_N^*(\alpha) := \begin{cases} N, & |\alpha| \leq 1/N \\ 1/|\alpha|, & 1/N \leq |\alpha| \leq \pi \end{cases}$$

and let $L_N : \mathbf{R} \to \mathbf{R}$ be the 2π-periodic extension of L_N^*. The following facts about L_N are known form [1]:

Lemma 1 *1.*

$$\exists K \forall N, \beta, \gamma \int_{-\pi}^{\pi} L_N(\beta - \alpha) L_N(\alpha - \gamma) d\alpha \leq K L_N(\beta - \gamma) \log N$$

2. If h is of bounded variation, then $\exists K$ such that $\forall N, s \leq N$ and $\forall \lambda$ we have $|H_s(\lambda)| \leq |H_N(\lambda)| \leq K L_N(\lambda)$.

The next lemma is easily proved by induction on N:

Lemma 2

$$\begin{aligned} H_N\left(h\left(\frac{\cdot}{N}\right)g(\cdot),\lambda\right) &= g(N-1)H_N(\lambda) - \\ & \sum_{s=0}^{N-1}\left(g(s)-g(s-1)\right)H_s(\lambda). \end{aligned}$$

Proof of theorem 1. We first write $d_N(u,\lambda)$ in a usefull form that makes it easy to prove the theorem. Using the representation 1 postulated in definition 1 we get

$$d_N(u,\lambda) = \sum_{s=0}^{N-1} h\left(\frac{s}{N}\right)\sum_{n\in M} A^0_{n,[uT]-N/2+s+1,T}\exp(-i\lambda s) \text{ a.s. }.$$

First we replace $A^0_{n,[uT]-N/2+s+1,T}$ by

$$A_n\left(\frac{[uT]-N/2+s+1}{T}\right)\exp\left(i\theta_{n,T}([uT]-N/2+s+1)\right)$$

and then $A_n\left(\frac{[uT]-N/2+s+1}{T}\right)$ by $A_n(u)$ to get

$$\begin{aligned} d_N(u,\lambda) &= \sum_{s=0}^{N-1} h\left(\frac{s}{N}\right)\sum_{n\in M} A_n(u)\exp\left(i\theta_{n,T}([uT]-N/2+s+1)\right) \\ & \exp(-i\lambda s) + R_1 + R_2 \text{ a.s. }. \end{aligned} \tag{4}$$

For the error terms we have

$$\begin{aligned} R_1 &:= \sum_{s=0}^{N-1} h\left(\frac{s}{N}\right)\sum_{n\in M}\Big(A^0_{n,[uT]-N/2+s+1,T} - \\ & A_n\left(\frac{[uT]-N/2+s+1}{T}\right)\exp\left(i\theta_{n,T}([uT]-N/2+s+1)\right)\Big) \\ & \exp(-i\lambda s), \end{aligned}$$

$|R_1| \leq O(N/T)$ a.s. by 2 of definition 1 and

$$\begin{aligned} R_2 &:= \sum_{s=0}^{N-1} h\left(\frac{s}{N}\right)\sum_{n\in M}\left(A_n\left(\frac{[uT]-N/2+s+1}{T}\right) - A_n(u)\right) \\ & \exp\left(i\theta_{n,T}([uT]-N/2+s+1)\right)\exp(-i\lambda s), \\ |R_2| &\leq O\left(\frac{N^2}{T}\right)\sum_{n\in M}\sup_u |A_n'(u)| \leq O\left(\frac{N^2}{T}\right) \text{ a.s.} \end{aligned}$$

by the mean value theorem.

The following considerations are the only place in the proof, where techniques are used that are not already known from the case of processes with evolutionary spectral density. In equation 4, we replace $\exp\left(i\theta_{n,T}([uT]-N/2+s+1)\right)$ by $\exp\left(i\theta_{n,T}([uT]-N/2+1)\right)\exp(i\eta_n(u)s)$. Now we can (a.s.) write $d_N(u,\lambda)$ as

$$\begin{aligned}&\sum_{n\in M} A_n(u)\exp\left(i\theta_{n,T}([uT]-N/2+1)\right)H_N\left(\lambda-\eta_n(u)\right)\\ &+R_1+R_2+R_3.\end{aligned}\tag{5}$$

Here

$$R_3 := \sum_{n\in M} A_n(u)\exp\left(i\theta_{n,T}([uT]-N/2+1)\right)R_4(n)$$

and

$$\begin{aligned}R_4(n) \quad &:= \quad \sum_{s=0}^{N-1} h\left(\frac{s}{N}\right)\exp\left(i(\eta_n(u)-\lambda)s\right)\\ &\qquad \{\exp\left(i\left(\theta_{n,T}([uT]-N/2+s+1)-\theta_{n,T}([uT]-N/2+1)\right.\right.\\ &\qquad \left.\left.-\eta_n(u)s\right)\right)-1\}\\ &= \quad H_N\left(h\left(\frac{\cdot}{N}\right)g(\cdot),\lambda-\eta_n(u)\right),\end{aligned}$$

where

$$\begin{aligned}g(s) \quad &:= \quad \exp\left\{i\left(\theta_{n,T}([uT]-N/2+s+1)-\theta_{n,T}([uT]-N/2+1)\right.\right.\\ &\qquad \left.\left.-\eta_n(u)s\right)\right\}-1.\end{aligned}$$

We want to use lemma 2 to find an upper bound for $|R_4(n)|$. Therefore we have to investigate g. $g(0)=0$ and for $s>0$ we have

$$\begin{aligned}g(s) \quad = \quad &\exp\left(i\sum_{k=0}^{s-1}\quad \{\theta_{n,T}([uT]-N/2+s+1-k)-\right.\\ &\qquad\qquad \left.\theta_{n,T}([uT]-N/2+s-k)-\eta_n(u)\}\right)\\ &-1.\end{aligned}$$

By the mean value theorem there exists a finite K such that $|g(s)|\leq$

$$\begin{aligned}K\left|\sum_{k=0}^{s-1}\{\right. \quad &\theta_{n,T}([uT]-N/2+s+1-k)-\theta_{n,T}([uT]-N/2+s-k)\\ &\left.-\eta_n\left(\tfrac{[uT]-N/2+s+1-k}{T}\right)+\left(\eta_n\left(\tfrac{[uT]-N/2+s+1-k}{T}\right)-\eta_n(u)\right)\}\right|.\end{aligned}$$

By 5 of definition 1 and the mean value theorem this is $\leq O(N/T)+O(N^2/T)$. Further, there exists a K such that

$$\begin{aligned}|g(s)-g(s-1)| \quad &\leq \quad K\left|\theta_{n,T}([uT]-N/2+s+1)-\theta_{n,T}([uT]-N/2+s)\right.\\ &\qquad \left.-\eta_n(u)\right|\\ &\leq \quad O\left(\frac{N}{T}\right).\end{aligned}$$

Hence $|R_4(n)| \leq L_N(\lambda - \eta_n(u))O(N^2/T)$ by lamma 2 and

$$|R_3| \leq L_N(\lambda - \eta_n(u))O(N^2/T) \sum_{n \in M} |A_n(u)|.$$

Further for the main term of $d_N(u, \lambda)$ we have

$$\begin{aligned} &\left| \sum_{n \in M} A_n(u) \exp\left(i\theta_{n,T}([uT] - N/2 + 1)\right) H_N\left(\lambda - \eta_n(u)\right) \right| \\ &\leq O(1) \sum_{n \in M} |A_n(u)| L_N(\lambda - \eta_n(u)). \end{aligned}$$

Using the representation (5) of $d_N(u, \lambda)$, we now turn to the proof of the theorem.

$$\begin{aligned} B_N(u, \lambda) &= \sum_{n \in M} |A_n(u)|^2 \int_{-\pi}^{\pi} \frac{|H_N(\lambda - \eta_n(u))|^2}{2\pi H_{2,N}(0)} \phi(\lambda) d\lambda \\ &\quad + R_5 + R_6 + R_7 \text{ a.s.} \end{aligned} \tag{6}$$

The leading error terms are

$$\begin{aligned} R_5 &:= \frac{1}{2\pi H_{2,N}(0)} \int_{-\pi}^{\pi} \overline{R_3} \sum_{n \in M} A_n(u) \exp\left(i\theta_{n,T}([uT] - N/2 + 1)\right) \\ &\quad H_N\left(\lambda - \eta_n(u)\right) \phi(\lambda) d\lambda \end{aligned} \tag{7}$$

and

$$\begin{aligned} R_6 &:= \frac{1}{2\pi H_{2,N}(0)} \int_{-\pi}^{\pi} \sum_{n \neq m \in M} A_n(u) \overline{A_m(u)} \\ &\quad \exp\left(i\left(\theta_{n,T}([uT] - N/2 + 1) - \theta_{m,T}([uT] - N/2 + 1)\right)\right) \\ &\quad H_N\left(\lambda - \eta_n(u)\right) H_N\left(\eta_m(u) - \lambda\right) \phi(\lambda) d\lambda. \end{aligned}$$

The other error terms have been put into R_7. They are of lower order or can be treated in the same way as R_5 and R_6.

$$\begin{aligned} |R_5| &\leq O\left(\frac{N^2}{T}\right) O\left(\frac{1}{N}\right) \sum_{n,m \in M} |A_n(u)||A_m(u)| \\ &\quad \int_{-\pi}^{\pi} L_N(\lambda - \eta_n(u)) L_N(\eta_m(u) - \lambda) d\lambda \\ &\leq O\left(\frac{N \log N}{T}\right) \sum_{n,m \in M} |A_n(u)||A_m(u)| L_N(\eta_m(u) - \eta_n(u)) \\ &\leq O\left(\frac{N^2 \log N}{T}\right) \text{ a.s.} \end{aligned}$$

$$\begin{aligned}|R_6| &\leq O\left(\frac{1}{N}\right)\sum_{n\neq m\in M}|A_n(u)||A_m(u)| \\ &\quad \int_{-\pi}^{\pi} L_N(\lambda-\eta_n(u))L_N(\eta_m(u)-\lambda)d\lambda \\ &\leq O\left(\frac{\log N}{N}\right)\sum_{n\neq m\in M}\frac{|A_n(u)||A_m(u)|}{|\eta_m(u)-\eta_n(u)|}.\end{aligned}$$

Since $\frac{|H_N(\lambda-\eta_n(u))|^2}{2\pi H_{2,N}(0)}$ is an approximate identity, this proves the first part of theorem 1. For the second part we use similar arguments to see that

$$\begin{aligned}\mathrm{Var}(R_6) &\leq O\left(\frac{(\log N)^2}{N^2}\right)\sum_{n\neq m, l\neq k\in M}\left|E(A_n(u)\overline{A_m}(u)\overline{A_l}(u)A_k(u)\right| \\ &\quad L_N(\eta_m(u)-\eta_n(u))L_N(\eta_l(u)-\eta_k(u)) \\ &\leq O\left(\frac{(\log N)^2}{N}\right)\sum_{n\neq m, l\neq k\in M}\frac{\left|E(A_n(u)\overline{A_m}(u)\overline{A_l}(u)A_k(u)\right|}{|\eta_m(u)-\eta_n(u)|}\end{aligned}$$

■

Remarks:

1. Equation 1 of assumption A1 is restrictive and essential. It excludes e.g., that $\eta_n(u)$ converges to $\eta_m(u)$ $(n \neq m)$ as say $u \to 1/2$ and $\eta_n(u) = \eta_m(u)$ for $u \geq 1/2$. This example is also excluded by equations 2 and 3 in theorem 1. If we want to allow for such examples we have to reformulate those equations. Equation 2 could be changed to

$$\forall u \in [0,1] \exists K < \infty \sum_{(n,m):\eta_n(u)\neq\eta_m(u)}\frac{|A_n(u)|\,|A_m(u)|}{|\eta_n(u)-\eta_m(u)|} \leq K \text{a.s.}$$

   and equation 3 similarly. Then $B_N(u,\lambda)$ converges (a.s. or in quadratic mean respectively) to

$$\begin{aligned}&B(u,\lambda) + \sum_{n\neq m:\eta_n(u)=\eta_m(u)}\Big(A_n(u)\overline{A_m}(u)\phi(\eta_n(u)) \\ &\lim_{T\to\infty}\exp\left(i\left(\theta_{n,T}([uT]-N/2+1)-\theta_{m,T}([uT]-N/2+1)\right)\right)\Big),\end{aligned}$$

   provided this limit exists. Even if it exists it is not real in general. The convergence is no longer uniform in u. This shows that the interaction of very closely adjacent spectral lines can cause a lot of trouble.

2. Theorem 1 can be extended to the case of mixed evolutionary spectra. Assume that

$$X_{t,T} = X^d_{t,T} + X^c_{t,T}$$

where $X^d_{t,T}$ has a discrete evolutionary spectrum $F^d(u,\lambda)$ and $X^c_{t,T}$ has evolutionary spectral density $f^c(u,\lambda)$. Then under A1 and the assumptions of theorem 1 on $X^d_{t,T}$ and assumption A.1 of [1] on $X^c_{t,T}$ we have

$$B_N(u,\lambda) \to \int_{-\pi}^{\pi} f^c(u,\lambda)\phi(\lambda)d\lambda + B(u,\lambda)$$

in probability as $T \to \infty$. The convergence is uniform in u.

3. The theory of discrete evolutionary spectra can be extended to allow for finitely many discontinuities in $A_n(u)$ and $\eta_n(u)$. Assume for simplicity that $A_n(u)$ and $\eta_n(u)$ have a single jump of finite hight at $u = u_0$ for some n, where u_0 is independent of n. Then $B_N(u,\lambda)$ still converges to $B(u,\lambda)$ for $u \neq u_0$. $B_N(u_0,\lambda)$ converges to

$$\frac{\int_0^{1/2} h^2(v)dv}{\int_0^1 h^2(v)dv} \sum_{n\in M} |A_n(u_0-)|^2 \phi(\eta_n(u_0-)) +$$
$$\frac{\int_{1/2}^{1} h^2(v)dv}{\int_0^1 h^2(v)dv} \sum_{n\in M} |A_n(u_0+)|^2 \phi(\eta_n(u_0+)),$$

a.s. or in quadratic mean, if a.s. there exists a K such that for every u

$$\sum_{n\neq m} \frac{|A_n(u-)\overline{A_m(u-)}|}{|\eta_m(u-)-\eta_n(u-)|} \leq K,$$

$$\sum_{n\neq m} \frac{|A_n(u-)\overline{A_m(u+)}|}{|\eta_m(u-)-\eta_n(u+)|} \leq K \quad \text{and}$$

$$\sum_{n\neq m} \frac{|A_n(u+)\overline{A_m(u+)}|}{|\eta_m(u+)-\eta_n(u+)|} \leq K$$

or if there exists a K such that for every u

$$\sum_{n\neq m, l\neq k} \frac{|E(A_n(u-)\overline{A_m(u-)A_l(u-)}A_k(u-))|}{|\eta_m(u-)-\eta_n(u-)|} \leq K,$$

$$\sum_{n\neq m, l\neq k} \frac{|E(A_n(u-)\overline{A_m(u+)A_l(u-)}A_k(u+))|}{|\eta_m(u+)-\eta_n(u-)|} \leq K \quad \text{and}$$

$$\sum_{n\neq m, l\neq k} \frac{|E(A_n(u+)\overline{A_m(u+)A_l(u+)}A_k(u+))|}{|\eta_m(u+)-\eta_n(u+)|} \leq K$$

respectively.

Remark 1 is immediate from the proof of theorem 1. The proof of remark 2 is more technical than that of theorem 1. In addition to the methods presented here, it uses techniques from the theory of evolutionary spectral densities. We omit it here.

The main idea in the proof of remark 3 is to (a.s.) write $d_N(u, \lambda)$ as

$$\begin{aligned}&\sum_{n\in M} A_n(u-)\exp(i\theta_{n,T}([ut]-N/2+1))H_{N/2}(\lambda-\eta_n(u-)) + \\ &\sum_{n\in M} A_n(u+)\exp(i\theta_{n,T}([ut]-N/2+1)) \\ &\Big(H_N(\lambda-\eta_n(u+)) - H_{N/2}(\lambda-\eta_n(u+))\Big) + R\end{aligned}$$

where R is of reduced order. The details are technical and we omit them here.

3 Application to Licklider's theory of pitch perception

In 1951 Licklider proposed a theory of pitch perception ([4]), that will be called *correlogram* in the sequel. Because of its high computational costs not many sounds could be analyzed at that time using this model. In the last years the interest in the correlogram grew again (s. e.g. [5, 10]) because the computational capabilities had increased drastically. Slaney and Lyon were able to compute it in real time for the first time ([10]).

Here we investigate a somewhat simplified version of this model on the basis of the theory of discrete evolutionary spectra. First Licklider's theory is described. Then we discuss the asymptotics of the correlogram and present an algorithm, that computes this simplified version of the model much faster than the algorithm used by Slaney and Lyon ([11, 10]). Finally a simple pitch estimator based on the correlgram is investigated. We analyze its asymptotic behavior and how it works on processes with discrete evolutionary spectra that are very similar to amplitude modulated sounds. We are especially interested in the effect of the shift of the pitch of the residue described by Schouten et al. in [8].

3.1 The correlgram

3.1.1 Informal description

When we hear a sound the soundwave has traveled through our outer ear to hit the eardrum. From there, the vibrations were transferred to the cochlea (or inner ear) by three small ossicles in the middle ear.

The inner ear is a bony snail-like structure. If we uncoil it, it becomes a long straight tube that is partitioned by the basiliar membrane, that extends almost the entire length of the cochlea. When the sound enters the inner ear, a traveling wave on the basiliar membrane is caused. The place, where this wave has its maximum amplitude depends on the frequency of the sound. Now the movement of the basiliar membrane causes the hair cells to release a chemical transmitter that generates nerve impulses in the auditory nerve. Because the movement of the basiliar membrane is different at different places depending on the frequency of the sound, different groups of hair cells are activated by different frequencies. The distribution of the energy of the sound among different frequencies is mapped to the distribution of haircell activities at different places in the cochlea.

These facts about hearing seem to be uncontroversial and more information may be found in textbooks such as [3]. Now Licklider's assertion is, that in the brain for every place in the cochlea or every group of hair cells an autocorrelation of the neural activity caused by that group is computed. This will become clearer as soon as we describe the correlgram mathematically.

3.1.2 The mathematical model

A model of the outer, middle and inner ear has been proposed in [9]. We use the linear part of it to define a simplified correlogram. For details see [9].

The effect of the outer and middle ear on the soundwave are described by a linear filter. So the incoming sound is filtered first.

Next, the mapping of the energy distribution among frequencies to the distributions of basiliar membrane movement at different places of the cochlear is modelled by a filterbank. The cochlea is partitioned into 86 sections. For each sections there is a linear bandpass filter in the filterbank. The frequency responses of the individual filters are rather broad and have one peek. They differ in the position of the peek and their bandwidth: The higher the frequency of the peek is, the broader is the bandwidth. The frequency responses overlap strongly. The output of the outer-middle-ear-filter is filtered in every filter of the filterbank separatly. So we get a vector of 86 time series.

While Slaney and Lyon model the strongly nonlinear effects of the haircells, we leave this step out.

Now, for every such time series, the (empirical) autocovariance function is computed.

Let $(c_{p,j})_j, (p = 1, \ldots, 86)$ be the impulse response of the filter (of the filterbank) corresponding to section p of the cochlea convolved with the impulse response of the outer-middle-ear-filter. Further let $(X_t)_{t=1,\ldots,T}$ be the digitized input sound. Then the p-th component of the output vector of the filterbank is

$$Y_{p,t} := \sum_{j=0}^{\infty} c_{p,j} X_{t-j}$$

and the correlgram can be written as

$$\mathrm{KOR}_T^*(\tau, p, u) := \int_{-\pi}^{\pi} I_N^{Y_p}(u, \lambda) \exp(i\lambda\tau) d\lambda$$

where $I_N^{Y_p}(u, \lambda)$ is the tapered periodogram of a segment of $Y_{p,t}$. In fact, this is exactly what the algorithm given by Slaney and Lyon does, if we use it to compute our simplified correlgram. The incoming sound is filtered in the time domain (using the difference equations that describe the filters), then the periodogram is computed and the result is subjected to an inverse fourrier transform. This algorithm takes a lot of computing time, since for every section of the cochlea, a periodogram and an inverse fourrier transform have to be computed.

If we could do the filtering in the frequency-domain, we would be much faster, since we would have to compute the periodogram and the inverse fourrier transform only once for every u. But since we use a segmentwise periodogram and we cannot expect $Y_{p,t}$ to be stationary, it is not clear that this will lead to the same result as the procedure given above. In the next subsection we will show, that in fact we can do the filtering in the frequency-domain if $X_t = X_{t,T}$ has a discrete evolutionary spectrum.

3.2 Linear filters and discrete evolutionary spectra

Let $X_{t,T}$ be a process with discrete evolutionary spectrum $F^X(u, \lambda)$ and $(c_j)_{j \in \mathbf{N}}$ be the impulse response of a linear filter. Assume that

$$\sum_{j=0}^{\infty} c_j z^j = k(z) = \frac{a(z)}{b(z)},$$

where a and b are polynomials with real coefficients and $b(z) \neq 0$ for every complex number z such that $|z| \leq 1$.

Theorem 2 *Then*

$$Y_{t,T} := \sum_{j=0}^{\infty} c_j X_{t-j,T}$$

can be written as

$$Y_{t,T} = \sum_{n \in M} A_n \left(\frac{t}{T}\right) \exp(i\theta_{n,T}(t))\, k\left(\exp\left(-i\eta_n\left(t/T\right)\right)\right) + R \; a.s. \;,$$

where $|R| \leq O\left(\frac{1}{T}\right)$ *a.s. . Hence* $Y_{t,T}$ *has spectral distribution function*

$$F^Y(u,\lambda) = \left|k\left(\exp\left(-i\eta_n\left(t/T\right)\right)\right)\right|^2 F^X(u,\lambda).$$

Proof

$$Y_{t,T} = \sum_{j=0}^{\infty} c_j \sum_{n\in M} A^0_{n,t-j,T} \text{ a.s. .}$$

Again we replace $A^0_{n,t-j,T}$ by $A_n\left(\frac{t-j}{T}\right)\exp(i\theta_{n,T}(t-j))$ and $A_n\left(\frac{t-j}{T}\right)$ by $A_n\left(\frac{t}{T}\right)$, making errors R_1 and R_2 with

$$\begin{aligned} |R_1| &\leq O\left(\frac{1}{T}\right)\sum_{j=0}^{\infty}|c_j| \text{ a.s.} \\ |R_2| &\leq O\left(\frac{1}{T}\right)\sum_{j=0}^{\infty}j|c_j| \text{ a.s.} \end{aligned}$$

Now $\sum_{j=0}^{\infty} jc_j = \frac{\partial k(z)}{\partial z}$ at $z=1$ and hence converges absolutely. We have

$$\begin{aligned} Y_{t,T} &= \sum_{n\in M} A_n\left(\frac{t}{T}\right)\exp(i\theta_{n,T}(t)) \\ &\quad \sum_{j=0}^{\infty} c_j \exp(i(\theta_{n,T}(t-j)-\theta_{n,T}(t))) + R_1 + R_2 \text{ a.s. .} \end{aligned}$$

Replacing $\theta_{n,T}(t-j)-\theta_{n,T}(t)$ by $j\eta_n\left(t/T\right)$ we get the result, making an error R_3 such that

$$|R_3| \leq O\left(\frac{1}{T}\right)\sum_{j=0}^{\infty} j^2|c_j| \text{ a.s. .}$$

But

$$\sum_{j=0}^{\infty} j^2 c_j = \frac{\partial k(z)}{\partial z} + \frac{\partial^2 k(z)}{(\partial z)^2}$$

at $z=1$ and therefore converges absolutely. ■

Remark: An analogous result holds for processes with mixed evolutionary spectra. The results cease to hold, if the spectrum has discontinuities in u.

3.3 The asymptotic behavior of the correlgram and a fast algorithm

From theorem 1 we see that if the input-sound has an evolutionary spectrum,

$$\text{KOR}_T(\tau,p,u) := \int_{-\pi}^{\pi} I_N^X\left(u,\lambda\right)\left|k_p\left(\exp\left(-i\lambda\right)\right)\right|^2 \exp(i\lambda\tau)d\lambda,$$

where $\left|k_p\left(\exp\left(-i\lambda\right)\right)\right|^2$ is the frequency response of filter $(c_{p,j})_j$ converges to the same quantity as $\text{KOR}^*_T(\tau,p,u)$ does.

Definition 3 *This quantity*

$$KOR(\tau, p, u) := \sum_{n \in M} |A_n(u)|^2 \left| k_p \left(\exp\left(-i\eta_n(u) \right) \right) \right|^2 \exp(i\eta_n(u)\tau)$$

is called theoretical linear correlogram.

We have:

Theorem 3 *Under assumption A1 and the assumptions of theorem 1, $KOR_T(\tau, p, u)$ and $KOR^*_T(\tau, p, u)$ both converge to $KOR(\tau, p, u)$ a.s. or in quadratic mean respectively.*

The definition of $\mathrm{KOR}_T(\tau, p, u)$ gives us an algorithm that is much faster than the one proposed by Slaney and Lyon. But note, that they aim at computing a nonlinear correlogram that can't be computed with the algorithm presented here.

3.4 Visualizing sounds with correlograms

The correlogram and hence also its input-sound may be visualized as a movie. The time u is represented by itself: $\mathrm{KOR}_T(\tau, p, u)$ is shown at time $[uT]$. For a fixed u $\mathrm{KOR}_T(\tau, p, u)$ is presented as a two-dimensional picture. For every τ and p we have one pixel. τ is plotted on the horizontal and p on the vertical axis. If $\mathrm{KOR}_T(\tau, p, u) < 0$ the pixel (τ, p) is red, else it is grey.[1] The bigger $|\mathrm{KOR}_T(\tau, p, u)|$ is, the darker is the pixel.

It turned out, that for most sounds a few cochlea-sections are so predominant, that the biggest part of the picture is white. Therefore the information contained in the correlogram is conveyed much better, if it is rescaled. We do the rescaling in exactly the same way as Slaney and Lyon:

$$\mathrm{KORR}_T(\tau, p, u) := \frac{\mathrm{KOR}_T(\tau, p, u)}{\mathrm{KOR}_T(0, p, u)^{0.75}}$$

The scaling with $\mathrm{KOR}_T(0, p, u)^{0.75}$ seems to be ad hoc. 0.75 is the exponent that made the correlogram look best. Note that we could not have used $\mathrm{KOR}_T(0, p, u)$ because then the differences between the cochlea-sections had been lost.

In [10] many correlgrams of interesting sounds are shown. Figure 1 presents a (rescaled) correlogram of the phoneme /A/[2] computed with the algorithm given in [11]. Since this is a stationary sound, all pictures look equal.

Figure 2 presents one computed with our algorithm:

[1] If you have not printed this paper on a color printer, you see the absolute value of the correlogram in figure 2. This paper is available as postscript file with color via anonymous ftp from statlab.uni-heidelberg.de.

[2] The transcription is according to the ARPAbet. See [2].

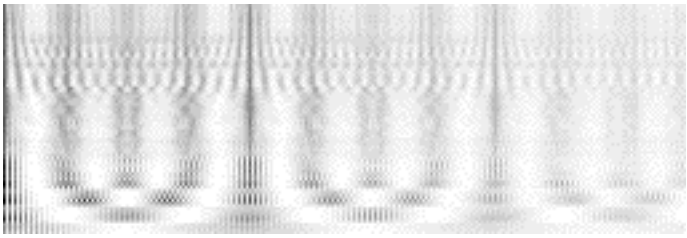

Figure 1: A correlogram of the phoneme /A/, computed with the algorithm given by Slaney and Lyon.

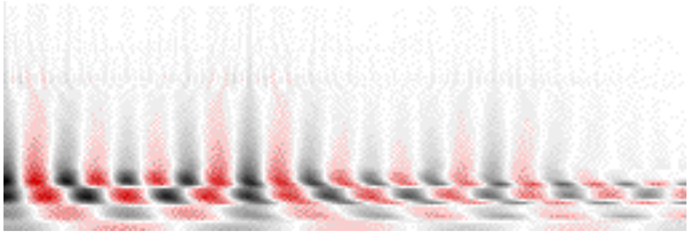

Figure 2: A correlogram of the phoneme /A/, computed with our algorithm. Note that if you don't have color, you only see a representation of the absolute value of the correlgram.

Differences along the vertical axis show us differences in the activity of different cochlea-sections and hence in the magnitude of energy in different frequency bands. Since the filters of the cochlea filterbank are tuned broadly, this gives us information about a strongly smoothed version (or the envelope) of the spectrum of the input-sound. Dark horizontal bands in the correlgram therefore indicate frequency bands with strong energy. In the context of speech-analysis they are often called *formants*.

In contrast, the correlation plotted on the horizontal axis reacts to the fine-structure of the spectrum of the input-sound. Assume e.g., the sound is a real valued locally stationary process, that has a discrete spectrum with lines at say η_1, η_{-1} and integer multiples. Then the correlation will be big at the lag corresponding to η_1, indicating the fundamental frequency of the sound. Therefore, dark vertical bands in the correlogram show pitch-information.

3.5 Pitch estimation

The last remark indicates that we can try to estimate the pitch of a sound by summing up the correlogram along the vertical axis (i.e., along the cochlea-sections) and looking for the maximum.

Definition 4

$$SUMKOR_T(\tau, u) := \sum_{p=1}^{86} KOR_T(\tau, p, u)$$

is called empirical summary correlgram.

$$PITCHPERIOD_T(u) := argmax_{\tau_1 \leq \tau \leq \tau_2} SUMKOR_T(\tau, u),$$

where τ_1 and τ_2 constitute some reasonable bound for the pitch period.

$$SUMKOR(\tau, u) := \sum_{p=1}^{86} KOR(\tau, p, u)$$

is called theoretical summary correlgram.

$$PITCHPERIOD(u) := argmax_{\tau_1 \leq \tau \leq \tau_2} SUMKOR(\tau, u).$$

Proposition 3 *If the assumptions of theorem 1 and A1 hold, $SUMKOR_T(\tau, u)$ converges to $SUMKOR(\tau, u)$ and $PITCHPERIOD_T(u)$ to $PITCHPERIOD(u)$ a.s. or in probability respectively.*

The proof of the first part of the proposition is trivial. The second part may be proved by arguments that are well known from consistency proofs for minimum distance estimators. An example of such a proof may be found in [1]. We therefore omit it.

Much more interesting than these theoretical results is the question, how good the pitch estimator describes real pitch perception by humans. A lot of psychopysical data about pitch perception is known. We want to test our pitch estimator against the observations about the pitch of the residue described by Schouten et al. in [8].

Schouten et al. presented amplitude modulated signals of the form

$$s(x) = 0.5m \sin(2\pi(f - g)x) + \sin(2\pi f x) + 0.5m \sin(2\pi(f + g)x)$$

to their listeners who judged the pitch of the signal by adjusting a matching signal. This matching signal was of the form

$$0.5m \sin(2\pi(n-1)\gamma g_0 x) + \sin(2\pi n \gamma g_0 x) + 0.5m \sin(2\pi(n-1)\gamma g_0 x), \qquad (8)$$

for some integer n, where γ was the parameter that could be adjusted by the listeners. The sound of interest was said to have pitch γg_0 for a subject, if the subject judged this sound to have the same pitch as the matching sound with parameter γ. Thus the pitch was given as a frequency in Hz. See [8] for more details. Schouten et al. used the values $m = 0.9$ and $g = g_0 = 200Hz$ and started with a value of $f = f_0 = ng_0$

where n is a natual number, typically $n = 10$. Then they shifted f up and down in steps of 50 Hz. The result was, that for $f = f_0$ the pitch was g_0. As f was shifted, the pitch changed linearly as long as f was close enough to f_0 i.e., $|f - f_0| < g_0$. A first approximation is

$$P = g_0 + \Delta f \frac{g_0}{f_0}, \text{ (for } |\Delta f| < g_0),$$

where P is the pitch (in Hz) and $\Delta f = f - f_0$. This is called the *first effect of pitch shift.* If one looks closer, one sees that the slope of the pitch as function of f is actually steeper. It can better be described as

$$P = g_0 + \Delta f \frac{g_0(1 + b)}{f_0}, \text{ (for } |\Delta f| < g_0),$$

where b depends on the indiviual subject that listens.[3] A typical value is $b = 0.35$. This result is called the *second effect of pitch shift.*

Now $s(x)$ is a deterministic signal and not a locally stationary process. Therefore we use a somewhat different but similar signal. Let

$$s_1(x) = 0.5m\cos(2\pi(f - g)x + \varphi_l') + \cos(2\pi f x + \varphi_c') + 0.5m\cos(2\pi(f + g)x + \varphi_r')$$

where $\varphi_l', \varphi_c', \varphi_r'$ are independent identically distributed random phases. If s is digitized at a sampling rate σ we may view it as a process with discrete evolutionary spectrum.

$$\begin{aligned} \widetilde{s_1}(t, T) \quad = \quad & \overline{A_l}\left(\frac{t}{T}\right)\exp(-i\theta_l(t)) + A_l\left(\frac{t}{T}\right)\exp(i\theta_l(t)) + \\ & \overline{A_c}\left(\frac{t}{T}\right)\exp(-i\theta_c(t)) + A_c\left(\frac{t}{T}\right)\exp(i\theta_c(t)) + \\ & \overline{A_r}\left(\frac{t}{T}\right)\exp(-i\theta_r(t)) + A_r\left(\frac{t}{T}\right)\exp(i\theta_r(t)) \end{aligned}$$

where

$$\begin{aligned} \eta_l(u) &= 2\pi(f - g)/\sigma \\ \eta_c(u) &= 2\pi f/\sigma \\ \eta_r(u) &= 2\pi(f + g)/\sigma \\ \theta_l(t) &= t\eta_l \\ \theta_c(t) &= t\eta_c \\ \theta_r(t) &= t\eta_r \\ A_l(u) &= 0.5\exp(i\varphi_l) \\ A_c(u) &= 0.225\exp(i\varphi_c) \\ A_r(u) &= 0.225\exp(i\varphi_r). \end{aligned}$$

[3] Here we do not consider a change of g as Schouten et al. did.

$\varphi_l, \varphi_c, \varphi_r$ are independent identically distributed according to the uniform distribution on $[-\pi, \pi]$. We use these processes with the values 1950, 2000, 2050, 2100, 2150, 2200, 2250 and 2300 Hz for f and 200 Hz for g.

In fact, the difference between these signals and those used by Schouten et al. is not significant: It is theoretically insignificant, because we can develop an asymptotic theory for almost periodic deterministic signals that is completely analogous to the theory of locally stationary processes with discrete spectra. Just let A and A^0 be deterministic, leave out 4 and replace almost sure convergence by normal convergence in definition 1. Then we can prove analogous results and the theoretical summary correlogram for $\widetilde{s_1}(t, T)$ is the same as for the analogous deterministic signal with $\varphi_l = \varphi_c = \varphi_r = 0$. The difference also seems to be practically insignificant, since for both signals the pitch estimates are exactly the same.

In addition we present a signal $\widetilde{s_2}(t, T)$, where the center frequency f is changed continuously from 1950 to 2200 Hz. Here

$$\begin{aligned}
\eta_l(u) &= 2\pi(f - g + u250\text{Hz})/\sigma \\
\eta_c(u) &= 2\pi(f + u250\text{Hz})/\sigma \\
\eta_r(u) &= 2\pi(f + g + u250\text{Hz})/\sigma \\
\theta_l(t) &= 2\pi t(f - g + \tfrac{t}{2T}250\text{Hz})/\sigma \\
\theta_c(t) &= 2\pi t(f + \tfrac{t}{2T}250\text{Hz})/\sigma \\
\theta_r(t) &= 2\pi t(f + g + \tfrac{t}{2T}250\text{Hz})/\sigma \\
A_l(u) &= 0.5\exp(i\varphi_l) \\
A_c(u) &= 0.225\exp(i\varphi_c) \\
A_r(u) &= 0.225\exp(i\varphi_r).
\end{aligned}$$

We analyze the sounds with the pitch estimator described above and with an even simpler one, that is just the argmax of the estimated covariance function of the signal:

$$\text{PITCHPERIOD2}_T(u) := \text{argmax}_{\tau_1 \le \tau \le \tau_2} \int_{-\pi}^{\pi} I_N(u, \lambda) \exp(i\lambda\tau) d\lambda.$$

It turns out that both estimators PITCHPERIOD_T and PITCHPERIOD2_T produce exactly the same result. Obviously the influence of the cochlear filters on pitch estimation is small in our examples.

The results for $\widetilde{s_1}$ are shown in table 1 and figure 3. In figure 3 the estimated pitches are shown as dots. The upper line is the line described by the first effect of pitch shift for $f_0 = 2000$Hz, the lower line for $f_0 = 2200$Hz. We have translated the pitch periods (lag τ) into frequncies in Hz (P) by the formula $P = \sigma/\tau$, where σ is the sampling rate of the signal. Hence P is the frequency of the pure oscillation (discretized with sampling rate σ) with period τ. A theoretical justification for this formula will be given below (p. 23). Further, for the relavant values of P, the pitch

f in Hz	pitch in Hz
1950	195.122
2000	200.000
2050	205.128
2100	210.526
2150	195.122
2200	200.000
2250	205.128
2300	207.792

Table 1: Estimated pitches of $\widetilde{s_1}$.

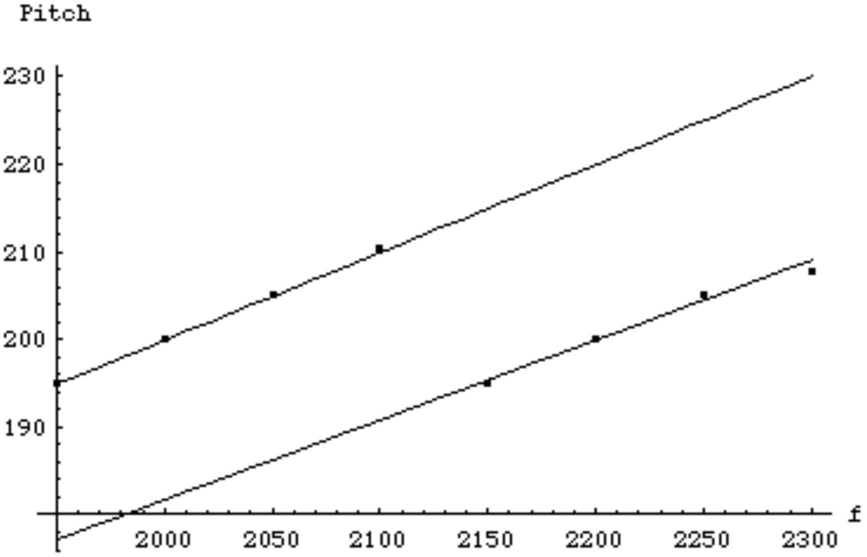

Figure 3: The estimated pitch of $\widetilde{s_1}$ vs. the center frequency plotted as dots. The lines show the first effect of pitch shift for $f_0 = 2000$Hz (upper line) and $f_0 = 2200$Hz (lower line).

estimate our estimators give for (a discretized version of) the matching signal (8) with $\gamma g_0 = P$, is the τ given by the above formula.

The first 4 pitch values almost lie on a line with a slope of approximatly g_0/f_0. Then a jump occurs and for the last 4 the slope is less steep. The first 4 values show almost exactly the behavior predicted by the first effect of pitch shift for $f_0 = 2000$Hz. But then the jump occurs too soon. The second 4 again show the behavior predicted by the first effect of pitch shift but not for $f_0 = 2000$Hz but for $f_0 = 2200$Hz instead. The second effect of pitch shift is not visible at all.

For the simple estimator PITCHPERIOD2$_T$ this behavior is also theoretically clear: PITCHPERIOD2$_T$ is the maximum of the empirical (tapered) covariance function of a segment of length N around uT of the input signal. By theorem 1 it converges to

$$c(u,\tau) := \sum_{n \in M} |A_n|^2 \exp(i\tau\eta_n(u)).$$

In our case, $c(u,\tau)$ is the theoretical covariance function and one can easily check, that it is an amplitude modulated pure oscillation.

$$\begin{aligned} c(u,\tau) &= 0.45\cos(\tau 2\pi(f-g)/\sigma) + \cos(\tau 2\pi f/\sigma) \\ &\quad +0.45\cos(\tau 2\pi(f+g)/\sigma) \\ &= (1 + 0.9\cos(2\pi\tau g/\sigma))\cos(2\pi\tau f/\sigma) \end{aligned}$$

with carrier frequency f and modulation frequency g. In fact, $c(u,\tau)$ is (up to a constant factor) the theoretical summary correlogram of $\widetilde{s_1}$, when all the linear filters are replaced by the identity transformation. Figure 4 shows it.

We call $\cos(2\pi\tau g/\sigma)$ the *modulating oscillation*, $1+0.9\cos(2\pi\tau g/\sigma)$ the *envelope* and $\cos(2\pi\tau f/\sigma)$ the *carrier oscillation* of $c(u,\tau)$.

Now if $f > g$, then $c(u,\tau)$ has it's maximum at that peek of the carrier oscillation that is the nearest one to a peek of the modulating oscillation or the envelope. The peeks of the carrier are at $\tau = k\sigma/f$, $k \in \mathbf{Z}$ and the peeks of the modulating oscillation and the envelope are at $\tau = l\sigma/g$, $l \in \mathbf{Z}$. Since the envelope of the empirical covariance function and also of the empirical summary correlogram falls off, as can be seen in figure 5, we take into account only $l = 1$. Now let us assume that we start with $f = f_0 = ng_0$ ($n \in \mathbf{N}$) and $g = g_0$. Then we clearly find the maximum of $c(u,\tau)$ at

$$\tau = \frac{\sigma}{g_0} = \frac{(f_0/g_0)\sigma}{f_0},$$

i.e., $k = f_0/g_0 = n$. If we now shift f upwards by an amount of Δf, still the "same" peek ($k = f_0/g_0$, $\tau = f_0\sigma/g_0(f_0+\Delta f)$) of the carrier will be the nearest to σ/g_0 as long as

$$\frac{\sigma}{g_0} - \frac{(f_0/g_0)\sigma}{f_0+\Delta f} < \frac{((f_0/g_0)+1)\,\sigma}{f_0+\Delta f} - \frac{\sigma}{g_0},$$

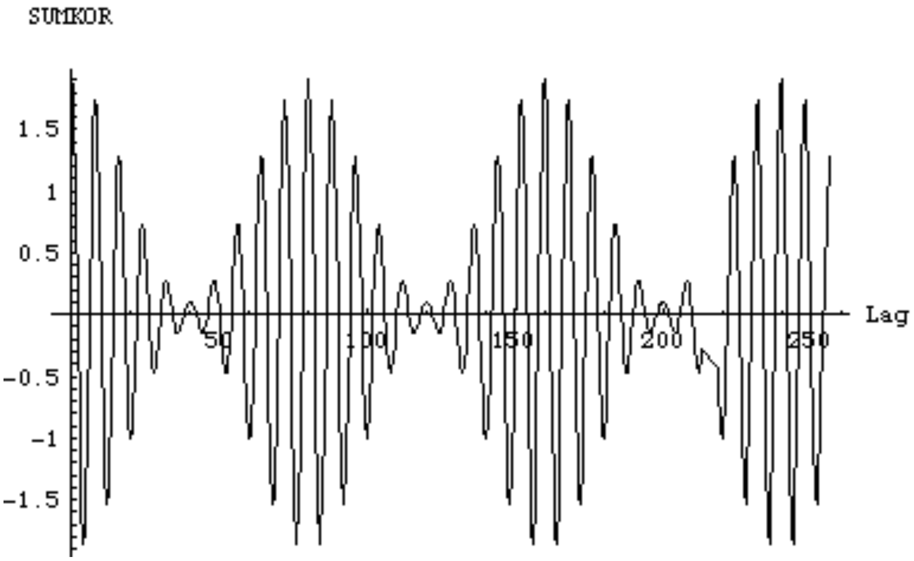

Figure 4: Theoretical covarianc function $c(u, \tau)$ of $\widetilde{s_1}$. The theoretical summary correlograms typically look very similar to this function. In fact, $c(u, \tau)$ is (up to a constant factor) the theoretical summary correlogram of $\widetilde{s_1}$, when all the linear filters are replaced by the identity transformation.

i.e., $\Delta f < g_0/2$. When Δf becomes $> g_0/2$, the argmax jumps to

$$\tau = \frac{((f_0/g_0) + 1)\,\sigma}{f_0 + \Delta f}$$

$(k = (f_0/g_0) + 1)$. Generally we see that

$$\text{argmax}_{\sigma/2g_0 \leq \tau \leq 3\sigma/2g_0} c(u, \tau) = \frac{\text{round}(f/g_0)\sigma}{f},$$

where

$$\text{round}(x) := \begin{cases} [x], & x - [x] < 0.5 \\ [x] + 1, & x - [x] > 0.5 \\ \text{undefined}, & x - [x] = 0.5 \end{cases}$$

Now these considerations also show, that a version of the matching signal (8) that is discretized at a sampling rate σ would be judged to have the pitch period

$$\tau = \frac{n\sigma}{n\gamma g_0} = \frac{\sigma}{\gamma g_0}$$

Hence PITCHPERIOD2 would judge, that $\widetilde{s_1}(t)$ has pitch $P = \gamma g_0$ if

$$\frac{\sigma}{\gamma g_0} = \frac{\text{round}(f/g_0)\sigma}{f}.$$

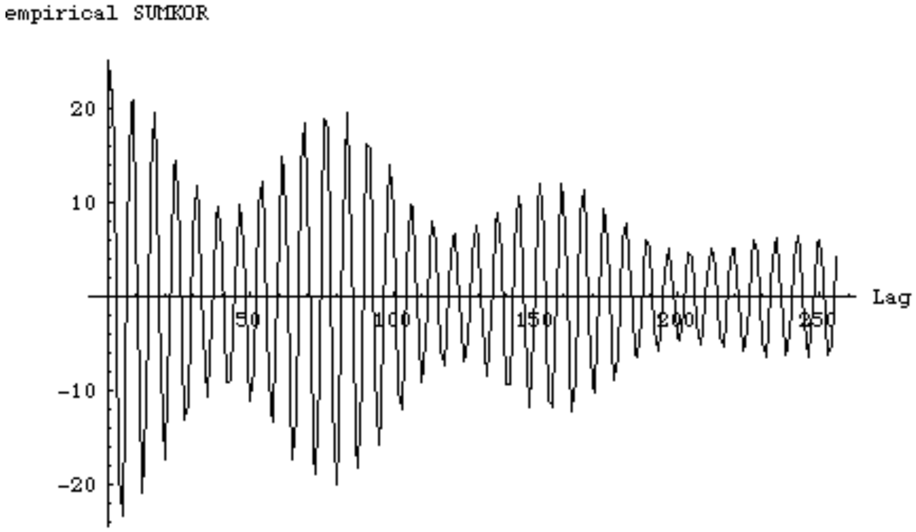

Figure 5: A typical empirical summary correlogram of $\widetilde{s_1}$.

This is equivalent to

$$P = \gamma g_0 = \frac{f}{\mathrm{round}(f/g_0)} = g_0 + \frac{(f - \mathrm{round}(f/g_0)g_0)}{\mathrm{round}(f/g_0)}$$

and for $f_0 := \mathrm{round}(f/g_0)g_0$ and $\Delta f = f - f_0$ this is the first effect of pitch shift

$$P = \gamma g_0 = g_0 + \Delta f \frac{g_0}{f_0}.$$

We also see why the estimated pitch jumps to a lower value at a carrier frequency of 2150 Hz. Then $f - 2000\,\mathrm{Hz} = 150\,\mathrm{Hz} > g_0/2 = 100\,\mathrm{Hz}$ and $\mathrm{round}(f/g_0)$ jumps to 11, while it was 10 as long as $f < 2100\,\mathrm{Hz}$.

For $\widetilde{s_2}$ the result looks very similar to that for $\widetilde{s_1}$. It is plotted in figure 6. Again the straight lines represent the first effect of pitch shift.

Summarizing our results we may say, that the pitch estimators analyzed here almost perfectly describe the first effect of pitch shift. The second effect is not visible at all. Further, since there is no mechanism that penalizes discontinuous behavior in $\mathrm{PITCHPERIOD}_T$ and $\mathrm{PITCHPERIOD2}_T$, the range near $|\Delta f| = g_0/2$, where the pitch is ambiguous according to the observations of Schouten et al. is much smaller (only one point) for our pitch estimators then for human listeners.

References

[1] Dahlhaus, R.: *Fitting time series models to nonstationary processes.* Preprint, Universität Heidelberg, 1992.

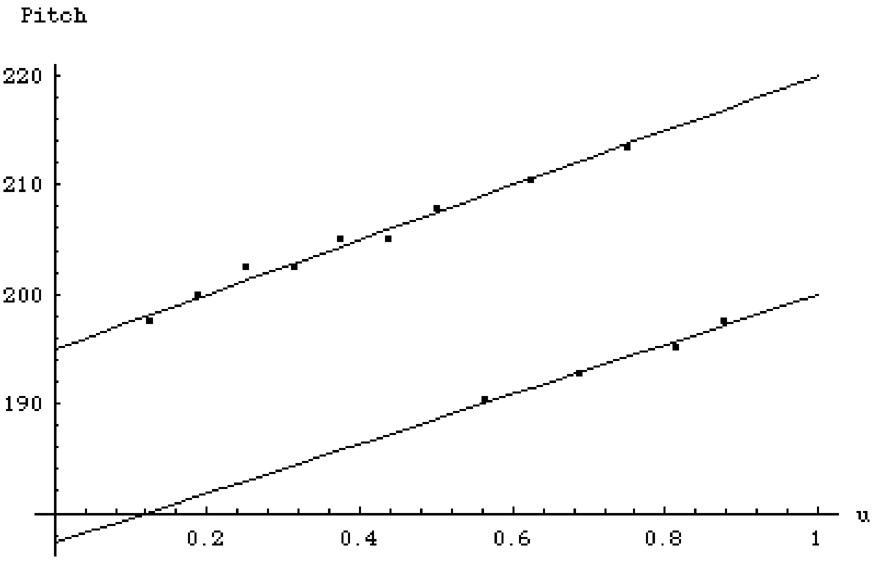

Figure 6: The estimated pitches of $\widetilde{s_2}$. Like in figure 3 the lines represent the first effect of pitch shift.

[2] Deller, J.R.jr., Proakis, J.G. & Hansen, J.H.L.: *Discrete-time processing of speech signals.* Macmillan Publishing Company, New York, 1993.

[3] Kelly, J.P.: *Hearing.* In Kandel, E.R., Schwartz, J.H. & Jessel, T.M. (eds): *Principles of neural science.* Elsevier science publishing, New York, 1991, 481–499.

[4] Licklider, J.C.R.: *A duplex theory of pitch perception.* Experientia 7, 1951, 128–134.

[5] Meddis, R. & Hewitt, M.J.: *Virtual pitch and phase sensitivity of a computer model of the auditory periphery I: Pitch identification.* J.Acoust.Soc.Am. 89, 1991, 2866–2882.

[6] Priestley, M.B.: *Spectral analysis and time series.* Vol 2, Academic Press, New York, 1981.

[7] Priestley, M.B.: *Non-linear and non-stationary time series analysis.* Academic Press, New York, 1988.

[8] Schouten, J.F., Ritsma, R.J. & Lopes Cardozo, B.: *Pitch of the residue.* J.Acoust.Soc.Am. Vol. 34, No. 8, 1962, 1418–1424.

[9] Slaney, M.: *Lyon's Cochlear Model.* Apple Technical Report ♯ 13, Apple Computer Inc., 1988.

[10] Slaney, M. & Lyon, R.F.: *Apple hearing demo real.* Apple Technical Report ♯ 25, Apple Computer Inc., 1991. (Available from the Corporate Library, Apple Computer, Cupertino, CA 95014.)

[11] Slaney, M. & Lyon, R.F.: *Visualizing soand with auditory correlograms.* unpublished, 1991.