Między czasem a częstością: elementy współczesnej analizy sygnałów.

9 marca 2004 roku

Oznaczenia

$s,\ s(t)$	sygnał, zwykle przedmiot analizy
$\hat{s}$	transformata Fouriera sygnału s
f	częstość $f = \frac{1}{T}$
ω	częstość kołowa $\omega = \frac{2\pi}{T}$
T	okres $\forall_t\ s(t+T) = s(t)$
$\equiv$	równe z definicji
$n!$	silnia: $3! = 1 * 2 * 3 = 6$
$\mathcal{F}$	przekształcenie Fouriera, $\mathcal{F}s(t) \equiv \hat{s}(\omega)$
i	$\sqrt{-1}$
$\overline{x}$	sprzężenie zespolone x: $\overline{a+ib} = a - ib$
	iloczyn (np. $2\pi = 6.283185...$)
$\langle \cdot , \cdot \rangle$	iloczyn skalarny, zwykle $\langle x(t), y(t) \rangle = \int x(t)\overline{y(t)}dt$ lub $\sum_i x_i \overline{y_i}$
$\star$	splot, $x(t) \star y(t) = \int x(\tau) y(t-\tau) d\tau$ lub $x \star y = \sum_i x_i y_{n-i}$
δ_k	delta Kroneckera: $\delta_k = 1$ dla $k = 0$, $\delta_k = 0$ dla $k \neq 0$
$\int$	brak granic całkowania oznacza $\int_{-\infty}^{\infty}$
$\forall x$	dla każdego x
$\exists x$	istnieje x
$\mathbb{Z}$	zbiór liczb całkowitych
$\mathbb{N}$	zbiór liczb naturalnych (włączając zero: $0 \in \mathbb{N}$)
$\mathbb{C}$	zbiór liczb zespolonych
$\mathbb{R}$	zbiór liczb rzeczywistych
$L^2(\mathbb{R})$	funkcje o skończonej energii $\int \|f(t)\|^2 dt < +\infty$
$L^2([a,b])$	funkcje o skończonej energii, określone na przedziale $[a,b]$: $\int_a^b \|f(t)\|^2 dt < +\infty$
$l^2(\mathbb{Z})$	dyskretne sekwencje o skończonej energii: $\sum \|f[n]\|^2 < +\infty$
$E(x)$	wartość oczekiwana x
$\|\|s\|\|$	$\ldots$

Spis treści

Rozdział 1

Wstęp

1.1 Sygnał

W języku potocznym sygnał to „wszelki umowny znak o treści informacyjnej”[13]. Z kolei w ramach tego wykładu przez „sygnał” rozumieć będziemy funkcję (zależną zwykle od czasu[1]) $s(t)$. Czy mówimy o tym samym?

Rozważmy przykładowe „znaki umowne”: czerwone światło na skrzyżowaniu to docierająca do naszych oczu fala elektromagnetyczna o długości ok. $\frac{0.76}{1000\,000 f}$ metra; znaczeniem tego sygnału jest „stój”. Litery, które czytasz, to pewien rozkład zaczernienia kartki, dający się niewątpliwie opisać z pomocą funkcji matematycznych, gdyż w tej właśnie formie przechodził kolejne etapy poprzedzające druk. Ich znaczenie ...

Wydaje się, że fizyczną postać informacji — umożliwiającą jej przekaz czy przechowywanie — można nazwać sygnałem. A odwrotnie? Czy każdy sygnał niesie ze sobą jakąś informację? Owszem, tylko niekiedy może ona być nieskończenie trudna do odczytania (dokładam wszelkich starań, aby nie miało to miejsca w przypadku tego tekstu :-). A sygnały zupełnie przypadkowe? Okazuje się, że nie są wcale powszechne ani łatwe do wytworzenia[2], więc właśnie całkowitą przypadkowość można uznać za niesioną przez sygnał informację.

1.2 Analiza

Informację niesioną przez milion liczb, wylosowanych niezależnie spomiędzy 0 i 1, można przedstawić krócej niż przez wyliczenie ich wszystkich — choćby tym właśnie zdaniem. Opis ten jest nie tylko bardziej zwięzły niż przytaczanie miliona wartości, ale oddaje jednocześnie najważniejsze ich cechy — *istotę* sygnału. Zwięzłych, trafny i kompletny opis sygnałów występujących w przyrodzie to właśnie Święty Graal analizy sygnałów. Ta książka to zaledwie zbiór wskazówek, które przy rozsądnym stosowaniu mogą nas czasem doprowadzić w jego pobliże.

Zastanówmy się więc, na czym właściwie ma polegać analiza czy opis sygnału, w przypadku bardziej skomplikowanym niż przytoczony powyżej? Sięgniemy raz jeszcze do *Słownika języka polskiego* [13]:

analiza(...)**1.** «myślowe, pojęciowe wyodrębnienie cech, części, lub składników badanego zjawiska lub przedmiotu; badanie cech elementów lub struktury czegoś oraz zachodzących między nimi związków (...)»

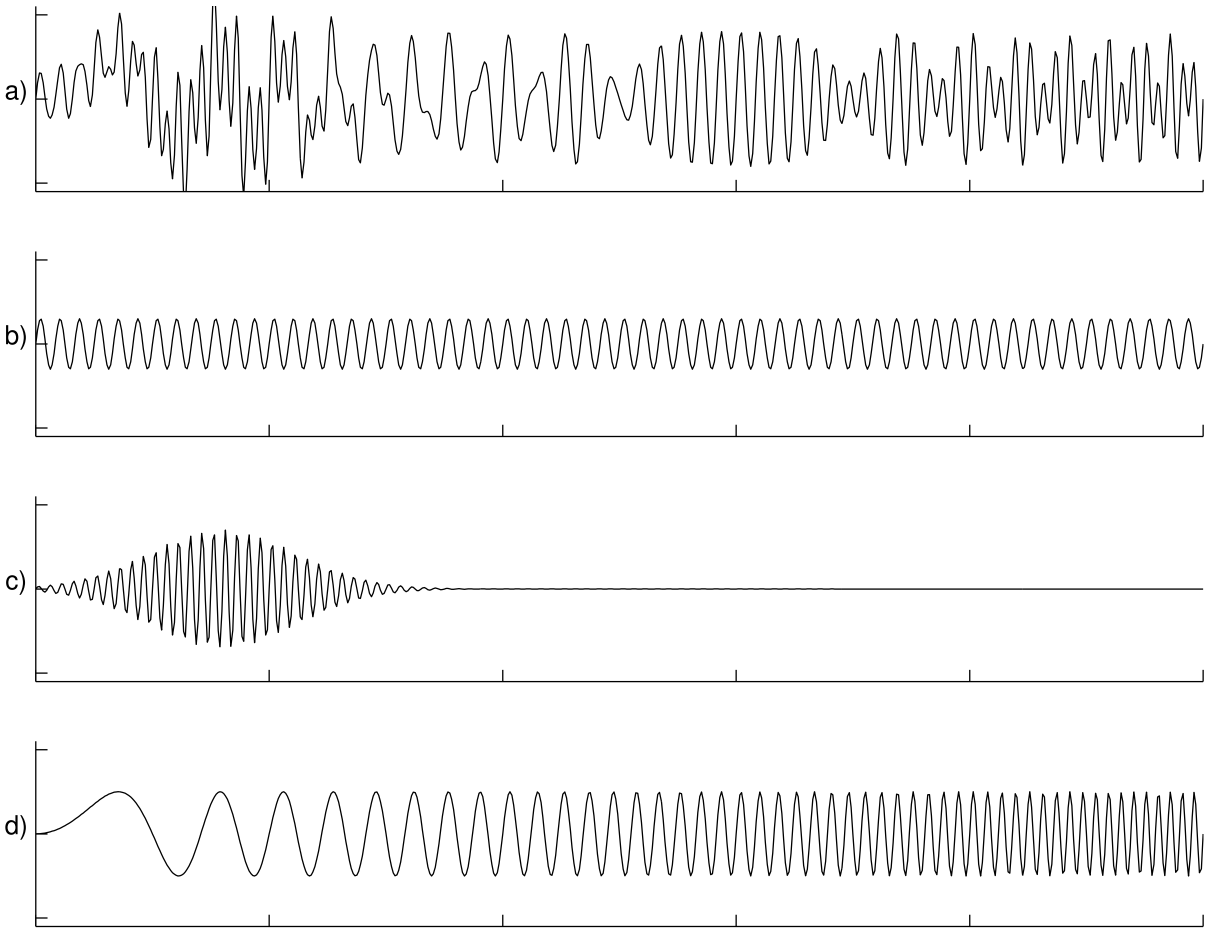
Skoncentrujmy się najpierw na «wyodrębnianiu części lub składników». Ilustrację tego podejścia stanowi rysunek 1.1.

„Tajemnicę” konstrukcji sygnału z górnej części rysunku 1.1 odkrywają wyrysowane pod nim funkcje składowe. Sygnał (a) jest ich (liniową) sumą. Taki przypadek sygnału będącego liniową kombinacją znanych funkcji możemy przedstawić ogólnie jako

---

[1] Uzyskane rezultaty nie zależą od fizycznej postaci zmiennej zależnej (t) i większość z nich stosowana jest np. w analizie obrazów.

[2] Za sygnał przypadkowy możemy uznać sekwencję liczb, przyjmującym wartości z określonego przedziału, np. od 0 do 1, z jednakowym prawdopodobieństwem. Ponadto w takiej sekwencji nie powinny występować *żadne* zależności między prawdopodobieństwem „wylosowania” następnej liczby a wartościami poprzednich, gdyż w nich właśnie może być zakodowana informacja. W przyrodzie znamiona takiej przypadkowości noszą zjawiska związane z rozpadem promieniotwórczym.



Rysunek 1.1: (a) = (b) + (c) + (d); $(b) = 0.3 \sin(2\pi 12t)$ (sinus); $(c) = 0.7e^{-(t-0.8)^2/0.2} \sin(2\pi 20t)$ (funkcja Gabora); $(d) = 0.5 \sin(2\pi\, 2\, t\, t)$ (chirp);

$$s(t) = \sum_k \alpha_k g_k \tag{1.1}$$

gdzie $\{g_i\}$ to zbiór „znanych" funkcji, a α_i to współczynniki określające ich wkłady. W konkretnym przypadku sygnału z rysunku 1.1 wyglądałyby one następująco:

$$\left\{ \begin{array}{ll} \alpha_1 = 0.3, & g_1 = \sin(2\pi 12t) \\ \alpha_2 = 0.7, & g_2 = e^{-(t-0.8)^2/0.2} \sin(2\pi 20t) \\ \alpha_3 = 0.5, & g_3 = \sin(2\pi\, 2\, t\, t) \end{array} \right. \tag{1.2}$$

Załóżmy, że interesująca nas w sygnale informacja została faktycznie zakodowana według równania (1.1). Niestety, dokładne „odgadnięcie" reprezentacji typu równania (1.2) jest w ogólnym przypadku — czyli w braku pewnej wiedzy a priori o sygnale — niemożliwe. Już sam wybór *rodzaju* funkcji (np. sinusy i kosinusy) jest nieskończenie trudny — wszak różnych funkcji jest nieskończenie wiele! Nawet gdy już zdecydujemy, jakiego rodzaju funkcje powinny najlepiej opisywać analizowany sygnał, to dobranie ich parametrów wciąż pozostaje poważnym problemem (patrz np. rozdział 4.2).

Ale analiza to również „badanie cech elementów lub struktury (...) oraz zachodzących między nimi związków". Możemy pokusić się o ustalenie związku między wartością sygnału w danej chwili i w chwilach poprzednich, w postaci zależności liniowej:

$$s(t) = \alpha_1 s(t - \Delta t) + \alpha_2 s(t - 2\Delta t) + \alpha_3 s(t - 3\Delta t) + ... \tag{1.3}$$

Jeśli weźmiemy pod uwagę czynniki przypadkowe, jak np. niedokładność pomiarów, do równań (1.1) i (1.3) należy dodać element stochastyczny — nie podlegający opisowi w ramach modelu szum ϵ (patrz rozdział 4):

$$s(t) = \sum_{k=0}^{M} \alpha_k g_k + \epsilon_M \tag{1.4}$$

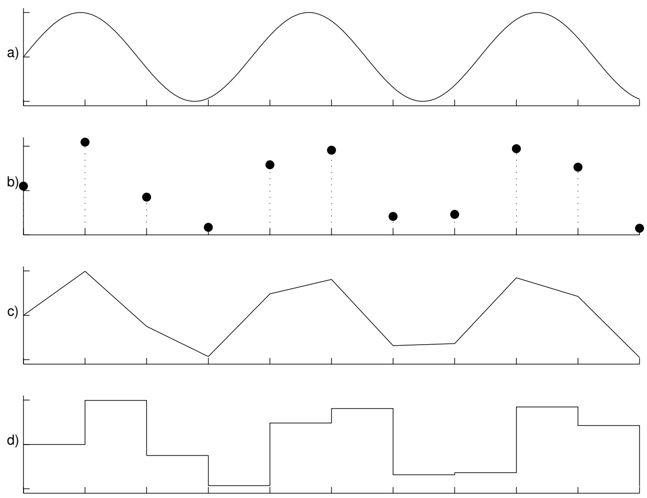
$$s(t) = \sum_{k=0}^{M} \alpha_k s(t - k\Delta t) + \epsilon_t \tag{1.5}$$

Na koniec zauważmy, że zaproponowane dotychczas modele mają postać *liniowych* sum. Uwzględnienie nieliniowości otwiera nowe, nie uwzględnione w tej książce rozdziały, jak np. chaos deterministyczny, fraktale ...

1.3 Sygnały ciągłe i dyskretne

Wartości akcji w chwilach zamknięcia kolejnych sesji giełdy tworzą sygnał z natury dyskretny. Jednak w przyrodzie większość stanowią sygnały ciągłe, jak dźwięk (zmiany ciśnienia powietrza w czasie) czy elektroencefalogram (EEG, potencjał elektryczny mózgu mierzony z powierzchni czaszki). Niezależnie od tego, współczesna analiza sygnałów odnosi się w praktyce głównie do sygnałów dyskretnych. „Winne” są oczywiście komputery, urządzenia z natury cyfrowe, czyli „rozumiejące” wyłącznie dyskretne wartości. Zastanówmy się nad wynikającymi stąd korzyściami i stratami.

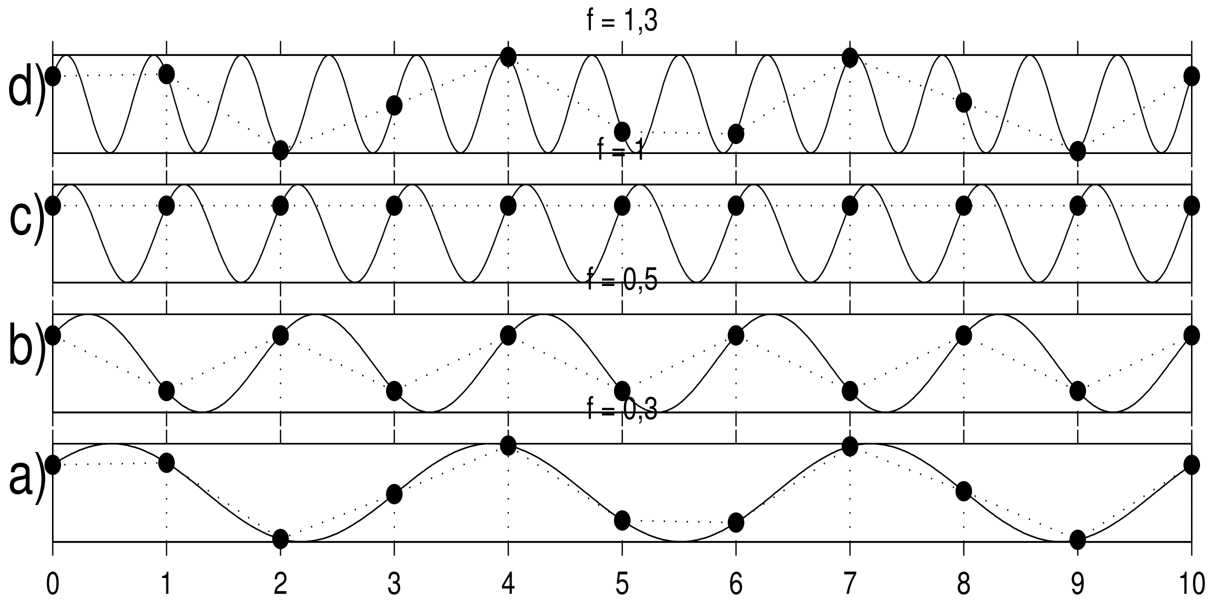
Jeśli sygnał z natury ciągły (np. dźwięk) zdecydujemy się analizować lub przechowywać w formie cyfrowej, to ciągłą funkcję (np. ciśnienia powietrza) w czasie musimy zastąpić jej wartościami zmierzonymi w określonych (najlepiej jednakowych) odstępach czasu, jak na rys. 1.2.



e) 1 1.992 0.751 0.07 1.482 1.81 0.315 0.363 1.844 1.426 0.049

Rysunek 1.2: Próbkowanie zamienia ciągły sygnał (a) na punkty (b) o współrzędnych w chwilach próbkowania i odpowiadających im wartościach sygnału ciągłego. Jeśli dysponujemy tylko sygnałem próbkowanym (b), to możemy „uzupełnić” wartości spomiędzy próbek przyjmując, że sygnał pomiędzy nimi jest np. liniowy (c) lub stały od poprzedniego punktu (d) — widzimy rozbieżności z sygnałem oryginalnym (a). Faktyczną reprezentacją funkcji po próbkowaniu jest ciąg liczb (e) plus znajomość odstępu próbkowania Δt. Optymalny sposób odzyskania wartości spomiędzy próbek, jeśli próbkowanie przeprowadzono zgodnie z regułami sztuki, podaje rozdział 2.6

Przy przejściu z reprezentacji ciągłej (rys. 1.2 a) do dyskretnej (b) tracimy informację o wartościach sygnału pomiędzy próbkami, a „naiwne” próby ich rekonstrukcji (c i d) znacznie odbiegają od oryginału (a).



Rysunek 1.3: Próbkowane z częstością 1 oscylacje o częstościach (f), od góry: 1,3, 1, 0,5 i 0,3. Sinusa o częstości 0,3 można odtworzyć dokładnie z samych wartości dyskretnych (kropki), podobnie dla granicznej częstości 0,5. Natomiast próbkowane z tą samą częstością szybsze oscylacje wprowadzają przekłamania — widoczna na samej górze oscylacja o częstości 1,3 daje w chwilach próbkowania wartości *dokładnie takie same* jak sygnał na dole. Zjawisko to nosi nazwę *aliasingu* (por. rozdział 2.5)

Pomimo tego, cyfrowy zapis dźwięku (płyty CD) zastąpił całkowicie analogowe „czarne płyty" z winylu — dlaczego?[3]

- Po pierwsze, przy pewnych dodatkowych założeniach o sygnale ciągłym $s(t)$, możliwe jest jego *dokładne* odtworzenie z dyskretnej sekwencji próbek, jeśli odstęp próbkowania Δt dobrano odpowiednio dla danego sygnału. Mowi o tym twierdzenie Nyquista (rozdział 2.6).
- Po drugie, zapis cyfrowy umożliwia *korekcję błędów*.

1.3.1 Zapis cyfrowy i korekcja błędów

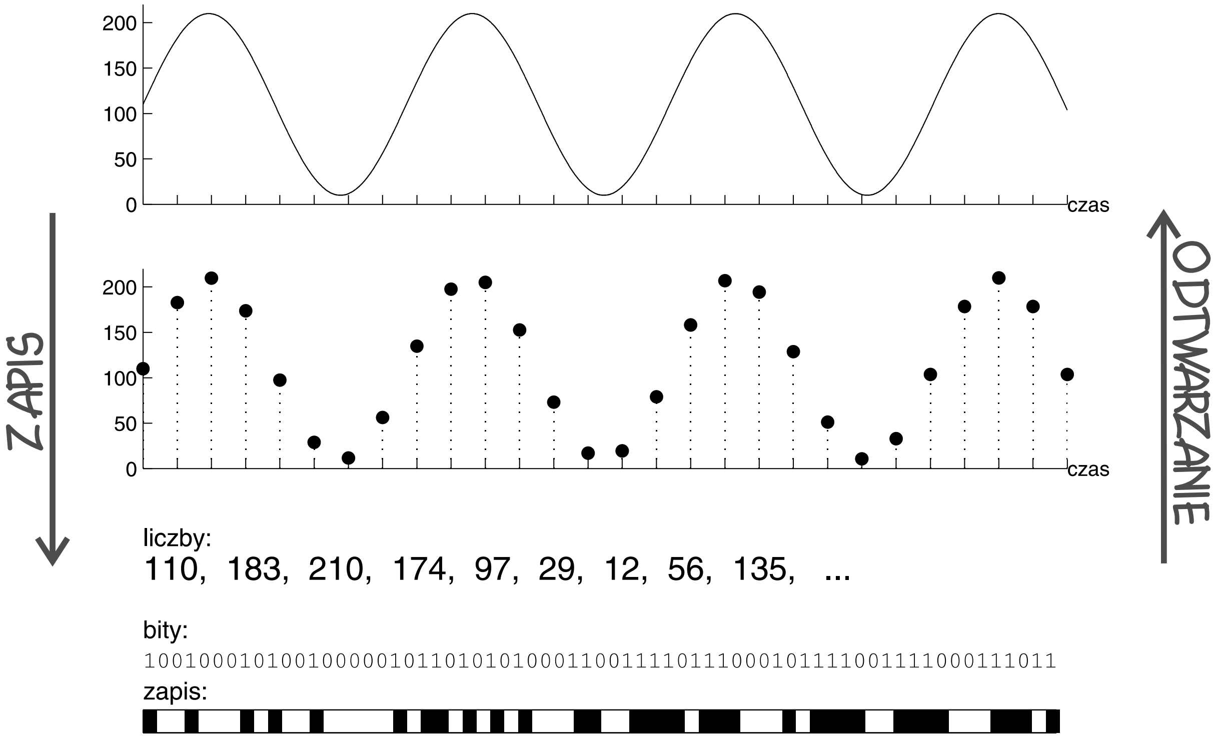
Aby zrozumieć, dlaczego łatwość korekcji błędów związana jest z zapisem cyfrowym, przyjrzyjmy się bliżej analogowym i cyfrowym zapisom dźwięku. Na płycie analogowej dźwięk kodowany jest w zmiennym wyżłobieniu rowka, w którym przemieszcza się igła gramofonu. W przybliżeniu możemy wyobrazić sobie, że „podskok" igły w większym wgłębieniu rowka odwzorowywany jest jako większe wychylenie membrany głośnika (po zamianie w impuls elektryczny i przejściu przez wzmacniacz). Tak więc wyżłobienie rowka płyty oryginalnie odwzorowuje dokładnie zapisany dźwięk. Jego zarysowanie lub zabrudzenie wprowadzi przy odtwarzaniu zakłócenia (zwykle trzaski). Jednoznaczne rozróżnienie, które z wyżłobień rowka winylowej płyty odzwierciedlają oryginalny zapis muzyki, a które powstały skutkiem uszkodzeń, jest właściwie niemożliwe, dlatego też muzyka ze starych płyt kojarzy nam się z obecnością trzasków i szumu.[4]

[3]Odpowiedź nie kryje się (niestety) w niższej cenie nośnika. Pomimo, że technologia cyfrowa faktycznie pozwala na znacznie tańszą produkcję (tj. powielanie) przy zachowaniu wysokiej jakości — jak wyjaśnimy za chwilę — to jednak cena średnio dwukrotnie wyższa niż cena odp. płyty winylowej, która w pierwszym okresie była uzasadniona wysokimi kosztami wprowadzania nowej technologii, po jej rozpowszechnieniu pozostała na wywindowanym poziomie, podwajając zyski wytwórni fonograficznych

[4]Tak naprawdę sprawa nie jest beznadziejna:

- część zakłócen pochodzi z zanieczyszczeń; w tym przypadku zwykle pomaga delikatne czyszczenie płyty.
- Do pozostałych zakłóceń, których nie da się usunąć mechanicznie, stosuje się potężną metodologię analizy sygnałów (będącą przedmiotem następnych rozdziałów), która pomaga zgadnąć, które dźwięki w zapisie mogą pochodzić z zakłóceń. Zwykle jednak nie da się usunąć dokładnie wszystkich zakłóceń bez naruszenia brzmienia oryginału.

W przypadku zapisu cyfrowego możemy w prosty sposób *wykryć* fakt wystąpienie zakłóceń. Wyobraźmy sobie, że zapisujemy muzykę jako szereg liczb, opisujących amplitudę fali dźwiękowej mierzoną w ustalonych odstępach czasu (rys. 1.4; dla płyty kompaktowej $\Delta t = \frac{1}{44100}$ sekundy). Ponieważ urządzenie, które będzie zamieniać ten zapis z powrotem na muzykę, i tak musi być swego rodzaju specjalizowanym komputerem (odtwarzaczem CD:), to do programu odtwarzającego możemy wprowadzić pewną modyfikację. Umówmy się dla przykładu, że z każdych dziesięciu kolejnych liczb, do zapisu muzyki będziemy wykorzystywać tylko dziewięć, a ostatnią będziemy dobierać tak, żeby suma kolejnych dziesięciu liczb zawsze wynosiła np. milion.



Rysunek 1.4: Od góry: ciągły (analogowy) zapis fali dźwiękowej, poniżej próbkowanie, czyli wybór chwil, w których ją mierzymy, dalej zamiana zmierzonych wartości na liczby i liczb na bity. Pasek na dole może być np. fragmentem ścieżki na płycie CD: białe pola (zera) odbijają światło lasera, a czarne (jedynki) nie.

Taki sposób zapisu wprowadza redundancję, czyli nadmiar informacji w zapisie, ponieważ przy prawidłowym odczycie wystarczyłoby znać dziewięć kolejnych liczb, aby wyznaczyć dziesiątą (jako milion minus suma pozostałych dziewięciu). Jednak jeśli wczytamy z takiego zapisu wszystkie liczby, i suma którejś dziesiątki okaże się inna niz milion, to mamy pewność, że w tym miejscu wystąpił błąd.[5] Taka informacja jest bardzo cenna:

- Jeśli *jesteśmy pewni*, że nagły skok amplitudy w kilku kolejnych próbkach jest wynikiem błędu zapisu, a nie efektem zamierzonym przez muzyka, to możemy ten skok „przemilczeć”, czyli np. zastąpić „popsute” próbki średnią wartością poprzednich.

- Możemy zwiększyć redundancję i zapisać dwie jednakowe kopie; jeśli uszkodzeniu ulegnie fragment pierwszej kopii, program może automatycznie sięgnąć do odpowiedniego fragmentu drugiej kopii[6].

[5] Ale poprawna suma nie daje gwarancji, że błędu nie ma. W jednej dziesiątce mogą wystąpić np. dwa jednakowe błędy o przeciwnych znakach i suma pozostanie niezmieniona. Dlatego sumy kontrolne liczy się w bardziej wyrafinowany sposób (np. CRC – *Cyclic Redundancy Check*)

[6] Prawdopodobieństwo wystąpienia uszkodzeń w tych samych fragmentach dwóch zapisów jest już bez porównania mniejsze niż pojedynczego uszkodzenia. Sposobem wprowadzania nadmiarowości, który minimalizuje prawdopodobieństwo wystąpienia takich pechowych przypadków, rządzi dość złożona matematyka z pogranicza statystyki, której nie będziemy tu omawiać. W każdym razie, dwie jednakowe kopie umieszczone jedna za drugą zwykle nie okazują się rozwiązaniem otymalnym.

- W przypadku transmisji przez modem, program może zażądać powtórnego przesłania uszkodzonego fragmentu.

⋆ ⋆ ⋆

Niezależnie od tych korzyści, jeśli chcemy analizować sygnały z pomocą komputera *(maszyny cyfrowej)*, i tak jesteśmy „skazani" na pracę z ich dyskretną formą.

Mimo tego, większość ogólnych twierdzeń będziemy rozważać w przestrzeni funkcji ciągłych — o ile nie tyczą się *explicite* efektów próbkowania. Teoria funkcji ciągłych jest asymptotycznie zgodna z wynikami dla sekwencji dyskretnych — dla odstępu próbkowania dążącego do zera. Jej rezultaty, prostsze pojęciowo i łatwiejsze do wyprowadzenia, są wystarczająco dokładne by wyjaśnić ogólne własności dyskretnych obliczeń.

W uzasadnionych przypadkach będziemy oczywiście dyskutować efekty próbkowania; w takich sytuacjach będziemy rozróżniać sygnał ciągły $s(t)$ od dyskretnej sekwencji $s[n]$.