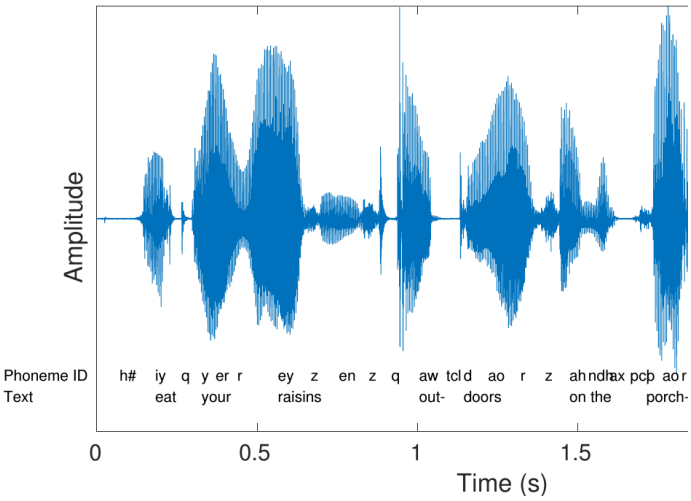
models

Waveform

由 Tom Bäckström创建, 最后修改于三月 28, 2019

Speech signals are sound signals, defined as pressure variations travelling through the air. These variations in pressure can be described as waves and correspondingly they are often called sound waves. In the current context, we are primarily interested in analysis and processing of such waveforms in digital systems. We will therefore always assume that the acoustic speech signals have been captured by a microphone and converted to a digital form.



A speech signal is then represented by a sequence of numbers x_n , which represent the relative air pressure at time-instant $n \in \mathbb{N}$. This representation is known as pulse code modulation often abbreviated as *PCM*. The accuracy of this representation is then specified by two factors; 1) the sampling frequency (the step in time between *n* and *n+1*) and 2) the accuracy and distribution of amplitudes of x_n.

Sampling rate

Sampling is a classic topic of signal processing. Here the most important aspect is the Nyquist frequency, which is half the sampling rate F_s and defines the upper end of the largest bandwidth $\left[0, \frac{F_s}{2}\right]$ which can be uniquely represented. In other words, if the sampling frequency would be 8000 Hz, then signals in the frequency range 0 to 4000 Hz can be uniquely described with this sampling frequency. The AD-converter would then have to contain a low-pass filter which removes any content above the Nyquist frequency.

The most important information in speech signals are the formants, which reside in the range 300 Hz to 3500 Hz, such that a lower limit for the sampling rate is around 7 or 8kHz. In fact, first digital speech codecs like the AMR-NB use a sampling rate of 8 kHz known as narrow-band. Some consonants, especially fricatives like /s/, however contain substantial energy above 4kHz, whereby narrow-band is not sufficient for high quality speech. Most energy however remains below 8kHz such that wide-band, that is, a sampling rate of 16 kHz is sufficient for most purposes. Super-wide band and full band further correspond, respectively, to sampling rates of 32 kHz and 44.1 kHz (or 48kHz). The latter is also the sampling rate used in compact discs (CDs). Such higher rates are useful when considering also non-speech signals like music and generic audio.

Accuracy and distribution of steps on the amplitude axis

In digital representations of a signal you are forced to use a finite number of steps to describe the amplitude. In practice, we must quantize the signal to some discrete levels.

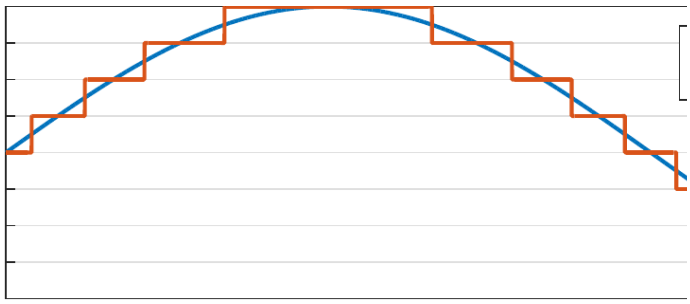
Linear quantization

models

Linear quantization with a step size Δq would correspond to defining the quantized signal as

$$\hat{x} = \Delta q \cdot \mathrm{round}(x/\Delta q).$$

The intermediate representation, $y = \mathrm{round}(x/\Delta q)$, can then be taken to represent, for example, signed 16-bit integers. Consequently, the quantization step size Δq has to be then chosen such that *y* remains in the range $y \in (-2^{15}, 2^{15}]$ to avoid numerical overflow.

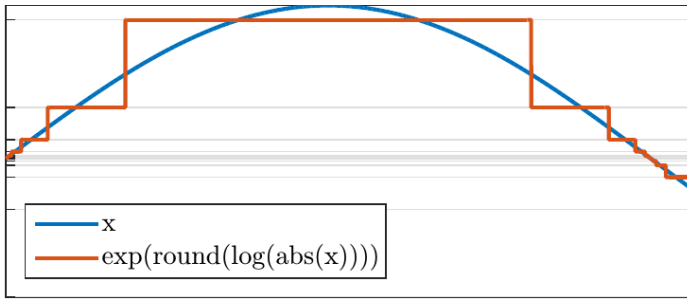


The beauty of this approach is that it is very simple to implement. The drawback is that this approach is sensitive to the choice of the quantization step size. To make use of the whole range and thus get best accuracy for *x*, we should choose the smallest Δq where we still remain within the bounds of integers. This is difficult because the amplitudes of speech signals vary on a large range.

Logarithmic quantization and mu-law

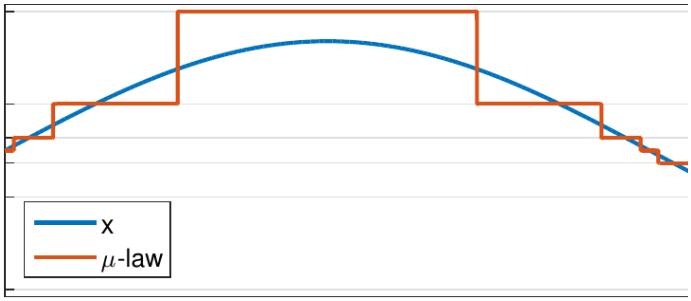
To retain equal accuracy for loud and weak signals, we *could* quantize on an logarithmic scale as

$$\hat{x} = \mathrm{sign}(x) \cdot \exp[\Delta q \cdot \mathrm{round}(\log(|x|)/\Delta q)].$$



Such operations which limit the detrimental effects of limited range are known as *companding* algorithms.

Here the intermediate representation is $y = \mathrm{round}(\log(|x|)/\Delta q)$ which can be reconstructed by $\hat{x} = \mathrm{sign}(x) \cdot \exp[\Delta q \cdot |y|]$. A benefit of this approach would be that we can encode signals on a much larger range and the quantization accuracy is relative to the signal magnitude. Unfortunately, very small values cause catastrophic problems. In particular, for *x=0*, the intermediate value goes to negative infinity $y = -\infty$, which is not realizable in finite digital systems.



A practical solution to this problem is quantization with the mu-law algorithm, which defines a modified logarithm as

$$F(x) := \mathrm{sign}(x) \cdot \frac{\log(1+\mu|x|)}{(1+\mu)}.$$

By replacing the logarithm with *F(x)*, we retain the properties of the logarithm for large *x*, but avoid the problems when *x* is small.

Wav-files

The most typical format for storing sound signals is the wav-file format. It is basically merely a way to store a time sequence, with typically either 16 or 32 bit accuracy, as integer, mu-law or float. Sampling rates can vary in a large range between 8 and 384 kHz. The files typically have no compression (no lossless nor lossy coding), such that recording hours of sound can require a lot of disk space. For example, an hour of mono (single channel) sound with a sampling rate of 44.1kHz requires 160 MB of disk space.