Speech Synthesis
a.k.a. TTS

Basic Components

- text processing
- text to phonetic/prosodic translation
- speech generation given phonetic and prosodic tags

Text Processing

- editing: “hte” → “the”
- acronyms: “a.k.a. TTS” → “also known as text to speech”
- abbreviations: “St.” can be “street” or “saint”
- numbers: “10” can be “one-zero” or “ten”
- symbols: $ (dollar), % (percent), @ (at)
- dates and times

Phonetic/Prosodic Translation from Text

- dictionary lookup or pronunciation rules
  determine the canonical pronunciation of each word to be synthesized, need to deal with complications such as liaisons and unknown words
- prosodic tag: patterns of duration, pitch, amplitude (hard problem)
- What is prodosy?
  the part of human communication that is not captured by the sequence of words

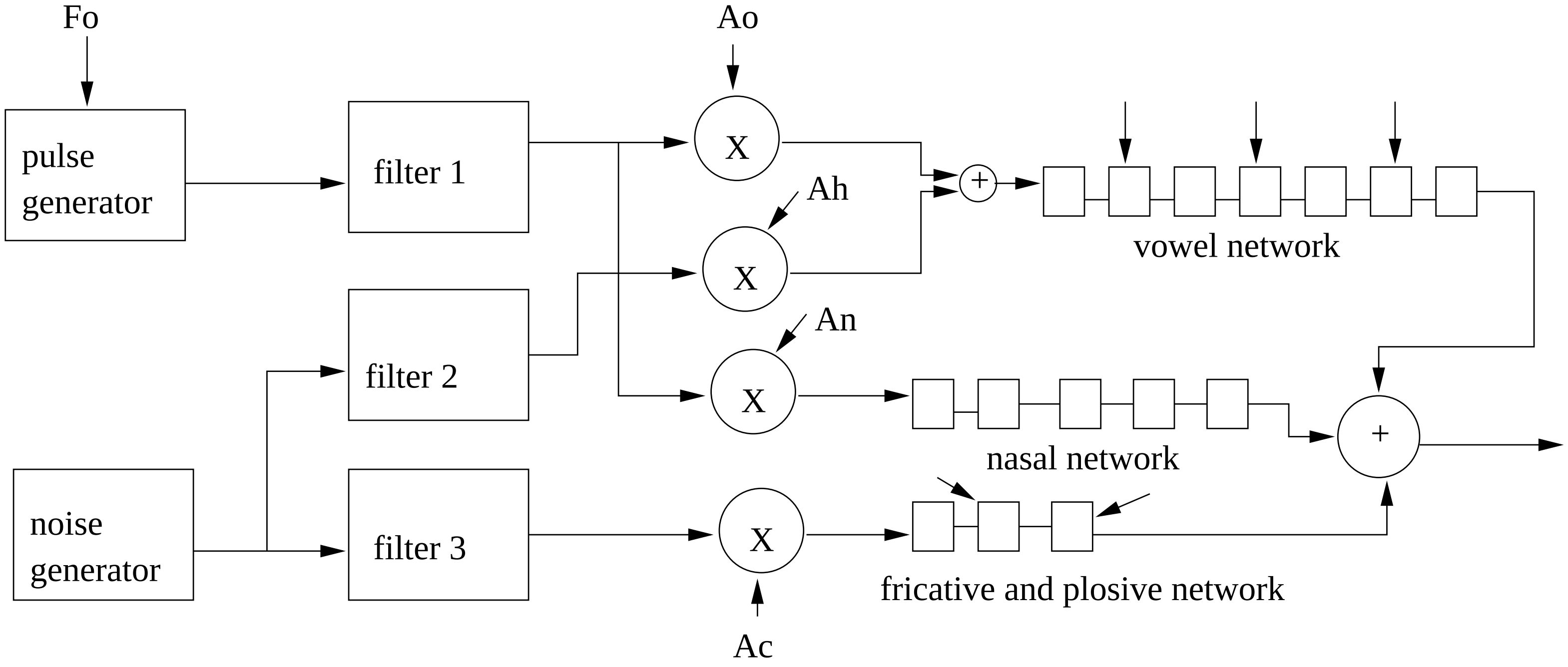
Speech Generation

- articulatory systhesis (quite outdated)
- source-filter synthesis

  use specific parameters to drive a system that enbodies the source-filter model of speech
- concatenative synthesis

  based on concatenation of stored templates of speech units, such as phones, syllables, words, etc.

Source Filter Synthesis



Concatenative Synthesis

- Units are collected from speech corpus via methods such as forced alignment.
- Each unit is labelled by its pitch, duration, amplitude and the identity of neighboring units.
- Given the prosody-tagged text, the database is searched for the optimal sequence of units.
- Algorithms such as dynamic programming are used for efficient search.
- There are two kinds of costs in the search process, the target cost and the concatenation cost. The design of cost functions is a research issue.