1 Introduction

The text-to-speech (TTS) synthesis procedure consists of two main phases. The first one is text analysis, where the input text is transcribed into a phonetic or some other linguistic representation, and the second one is the generation of speech waveform, where the acoustic output is produced from this phonetic and prosodic information. These two phases are usually called as high- and low-level synthesis. Procedure is presented in Figure 1.

Objectives of the sound synthesis are:

- maximize sound quality
- minimize memory usage
- minimize algorithm complexity
- minimize computational time
2 Speech Analysis

2.1 Source-Filter Model of Speech

- Models only the sound source and the formant frequencies, not any physical characteristics of the vocal tract.

- Linear model can be expressed as $S(z) = U(z)H(z)$, where $S(z)$ is speech, $U(z)$ is excitation and $H(z)$ transform function $Z$-transforms.

- Transfer function $H(z)$ can be approximated with all-pole filter derived in [2], thus the filter coefficients can be computed using linear prediction AR-model.

- Block diagram shown in Figure 2.

3 History

- Acoustical-mechanical synthesis
  - 1779 Kratzenstein: He explained physiological differences between long vowels and made apparatus to produce them artificially. Resonators are show in Figure 4.
  - 1791 Wolfgang Won Kempelen: “Vocal tract is the main source of the acoustic articulation.” He build a machine, which was able to produce single sounds and some sound combinations. One of his machine is show in Figure 5.
  - 1838 Willis: Connection between a specific vowel and geometry of the vocal tract.

- Electrical synthesis
Figure 3: Some milestones in speech synthesis.

- **1922** Stewart: His synthesizer had a buzzer as excitation and two resonant circuits to model the acoustic resonances of the vocal tract. The machine was able to generate single static vowel sounds with two lowest formants.

- **1939** Homer Dudley: VODER, Voice operating demonstrator. Considered as a first speech synthesizer. Inspired by vocoder by Bell Laboratories. The source signal was routed through ten bandpass filters whose output levels were controlled by fingers. Block diagram about VODER is show in Figure 6.

- **1953 Formant synthesizers**
  - **a)** Walter Lewrance: PAT, Parametric Artificial Talker (1953), which consisted of three electronic formant resonators connected in parallel. The input signal was either a buzz or noise.
  - **b)** Gunnar Fant: OVE, Orator Verbis Electris (1952), which consisted of formant resonators connected in cascade

- **1958** George Rosen, MIT: First **articulatory** synthesizer DAVO (Dynamic Analog of the VOcal tract) was controlled by tape recording of control signals created by hand.

- **1968** Noriko Umeda: First full text-to-speech system was based on an articulatory model and included a syntactic analysis module with sophisticated heuristics.

- **1976** Kurzweil: The first reading aid with optical scanner was introduced. It was used in libraries and service centers for visually impaired people

- Some milestones in history of speech synthesis are show in Figure 3.

## 4 Low level synthesis

- **System-modeling (Articulatory) synthesis** Tries to model human vocal organs as perfectly as possible, so it is potentially the most satisfying method to produce high-quality synthetic speech. On the other hand it’s very difficult to implement and it is computationally complex

- **Signal-modeling synthesis** Attempt to model the resulting signal.
  - Waveform synthesis
    - Concatenation
Figure 4: Kratzenstein’s resonators.

Figure 5: Von Kempelen’s third talking machine.
* Linear prediction  
* Sinusoidal models  
  - Spectrum synthesis  
    * Filterbank synthesis  
    * Formant synthesis

### 4.1 Articulatory synthesis

- Consist of models of human articulators and vocal cords. Human articulators is shown in Figure 7.
- Model parameters are updated toward target position for each phoneme using rules.
- Rules are set of area functions for each phonetic segment e.g. lip aperture, lip protrusion, tongue tip height...
- Data obtained from x-ray analysis. Problem: X-ray analysis in 2D ≠ Real world in 3D

### 4.2 Formant synthesis

- Model based on the source-filter-model of speech. The same idea used in VODER synthesizer in 1939.
- Formant synthesis provides infinite number of sounds which makes it more flexible than for example concatenation methods
Figure 7: Vocal tract outline with key parameters. C: tongue body center; H: hyoid; J: jaw; L: lips; T: tongue tip; V: velum

Figure 8: Basic structure of cascade formant synthesizer.

- At least three formants are generally required to produce intelligible speech and up to five formants to produce high quality speech. Each formant is usually modeled with a two-pole resonator which enables both the formant frequency (pole-pair frequency) and its bandwidth to be specified. ¹

- Synthesizers are usually controlled by rules, which determine which allophones are used in a certain context, and specify how these allophones and the transitions between them, should be produced.

- Can be divided further to cascade and parallel synthesis.

Cascade formant synthesis A cascade formant synthesizer shown in Figure 8 consists of band-pass resonators connected in series and the output of each formant resonator is applied to the input of the following one. The cascade structure needs only formant frequencies as control information. The main advantage of the cascade structure is that the relative formant amplitudes for vowels do not need individual controls. The cascade structure has been found better for non-nasal voiced sounds and because it needs less control information than parallel structure, it is then simpler to implement. However, with cascade model the generation of fricatives and plosive bursts is a problem.

Parallel formant synthesis A parallel formant synthesizer shown in Figure 9 consists of resonators connected in parallel. The excitation signal is applied to all formants simultaneously and their outputs are summed. Adjacent outputs of formant resonators must be summed in opposite phase to avoid unwanted zeros or antiresonances in the frequency response. The parallel structure enables controlling of bandwidth and gain for each formant individually and thus needs also more control information.

The parallel structure has been found to be better for nasals, fricatives, and stop-consonants, but some vowels can not be modeled with parallel formant synthesizer as well as with the cascade one.

\[ f = \frac{F_s}{2\pi} \quad \Delta f_{3dB} = -P\frac{\ln r}{2\pi} \]
4.3 Linear Prediction Synthesis

- Based on source-filter model.
- Originally developed to speech coding purposes.
- Excitation signal acts as glottal source and all-pole lp-filter mimics the vocal tract.
- Synthesis control parameters are either generated by a certain rules or estimated from the natural speech. Control parameters can be for example: fundamental frequency, gain i.e. frame energy and LP-coefficients.
- The LP-filter order is typically between 10 and 12 at 8 kHz sampling rate, but for higher quality at 22 kHz sampling rate, the order needed is between 20 and 24. The coefficients are usually updated every 5-10 ms.
- Warped Linear Prediction (WLP) takes advantages of human hearing properties and the needed order of filter is then reduced significantly, from orders 20-24 to 10-14 with 22 kHz sampling rate WLP provides better frequency resolution at low frequencies and worse at high frequencies. However, this is very similar to human hearing properties.
- Several other variations of linear prediction have been developed: Multipulse linear prediction (MLPC) where the complex excitation is constructed from a set of several pulses, Residual excited linear prediction (RELP) where the error signal or residual is used as an excitation signal and code excited linear prediction (CELP) where a finite number of excitations used are stored in a finite codebook.

4.4 Concatenation synthesis

- Operates by concatenating appropriate synthesis units to construct the required speech.
- synthesis units could be: words, syllables, diphones or monophones.
- Problem: How to extract the optimal collection of prototypes from natural speech and the developing of rules for concatenating them?

4.4.1 PSOLA-algorithm (Pitch-Synchronous Overlap and Add)

There are several versions of the PSOLA algorithm and all of them work in essence the same way. Time-domain version, TD-PSOLA, is the most commonly used due to its computational efficiency. The basic algorithm consist of three steps:

1. The analysis step where the original speech signal is first divided into separate but often overlapping short-term analysis signals (ST). Frame-length depends on the pitch.

2. The modification step of each analysis signal to synthesis signal.

3. The synthesis step where these segments are recombined by means of overlap-adding.

Another variations of PSOLA, Frequency Domain PSOLA (FD-PSOLA) and the Linear-Predictive PSOLA (LP-PSOLA), are theoretically more appropriate approaches for pitch-scale modifications because they provide independent control over the spectral envelope of the synthesis signal.

4.5 Sinusoidal Models

- Sinusoidal models are based on a well known assumption that the speech signal can be represented as a sum of sine waves with time-varying amplitudes and frequencies.

\[ s(n) = \sum_k A_k \cos(\omega_k n + \phi_k) \]

- Sinusoidal analysis/synthesis is shown in Figure 10

- While the sinusoidal models are perhaps very suitable for representing periodic signals, such as vowels and voiced consonants, the representation of unvoiced speech becomes problematic.

- Sinusoidal models are also used successfully in singing voice synthesis.

5 High level synthesis

- Text Preprocessing The input data should be converted to proper form for a synthesizer. For example if the character M is in some context converted as mega, the abbreviation MTV should not be converted as megaTV.

- Pronunciation Analysis for correct pronunciation from written text has also been one of the most challenging tasks in speech synthesis field. One way is to collect lexicon table for every word or try to develop a rule-based system.

- Prosody Prosodic consist of pitch, duration, and stress over the time. Almost everything seems to have effect on prosodic features of natural speech which makes accurate modeling very difficult
6 Exercise work

In this exercise work the linear prediction synthesis is introduced. Load data http://www.cs.tut.fi/~andip/kosta_aa.mat to matlab. Now you can find waveform of phoneme /a/ in variable y sampled at frequency found in variable Fs.

You should estimate the vocal tract filter parameters with LP-analysis (help lpc) and then use source-filter model with voiced sound and filter the periodic excitation with the vocal tract filter to synthesize vowel a.

You must pay attention on choosing the fundamental frequency of the excitation and order of the vocal tract filter. Excitation can be estimated as an impulse train.

Return commented matlab-code to andip@cs.tut.fi.

References

