Performing analysis and mapping

We extract energy, fundamental frequency and vibrato parameters. Energy and F0 curves are smoothed since micro-variations are already present in the SVS’s database samples. For the vibrato, we can either use the input performance curve or a vibrato template from an actual pre-recorded vibrato of the singer.

Phonetic Alignment:

Phonetic timing is derived by input performance. In offline mode, it uses the Julius [6] alignment module. Due to the quality loss when time-scaling unvoiced phonemes, two alignment alternatives are proposed: “voiced onset” and “vowel onset”.

Synthesis control layer:

Control parameters are: pitch, dynamics, phonetic timing and vibrato rate and depth. When building an internal score, the synthesizer selects the sample in the database with least distance cost to the control parameters.

> approaches for controlling expression

Control with expression models:

It deals with characterizing and modelling expressive performances to give expression to a neutral performance[5]. Current approaches work at symbolic or at signal level. User is limited to very high-level controls.

Control with input interfaces:

Musical controllers are employed in real-time, thus suitable for performance situations. In SVS, the problem is that there is no musical controller we know appropriate for singing voice.

Performance-driven control:

The control derives from an audio signal (Indirect Acquisition)[4] Here the mapping can be direct since voice is controlling voice.

> preliminary real-time implementation

It uses a HMM for phoneme recognition with a Japanese acoustic model of 16 gaussian mixtures. Text alignment is performed with a Finite State Network and a Viterbi algorithm. Current recognition latency is 5 frames (50ms) Concatenation process of diphone samples in the SVS introduces additional latency.

> main references: