



Singing Voice Resynthesis using Concatenation-based Techniques

Nuno Miguel da Costa Santos Fonseca

Goal

- Develop, create or adapt techniques for singing resynthesis
 - User's voice to control singing synthesis
 - Automatically recreate the same music and lyrics performance
- Using Sound and Music Computing (SMC) approaches:
 - Merging speech and music research
 - Sampling-based synthesis



Applications

- Audio Effect
 - Using an FX unit as a synthesizer, choosing the output sound
- New UI approach for Singing Synthesis
 - Using the user's voice to control a singing synthesizer
- Several contexts:
 - Restoration
 - Transynthesis
 - Instrument Enhancer



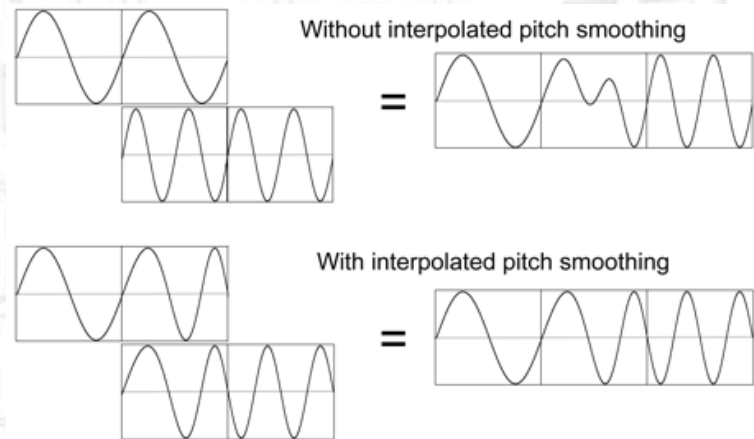
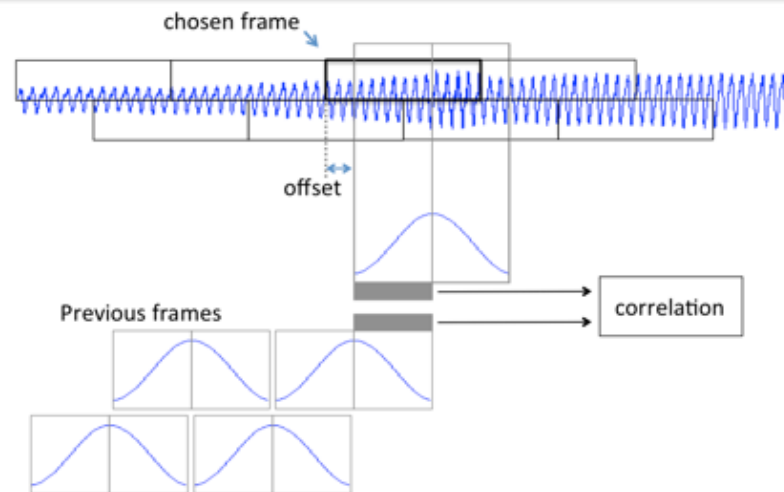
Singing Resynthesis

- Need to refocus on main goal (singing voice)
- Replication of three domains:
 - Phonetics
 - Pitch
 - Melodic Line
 - Pitch related effects (e.g. portamento, vibrato)
 - Dynamics (sound intensity)
 - e.g. Crescendo



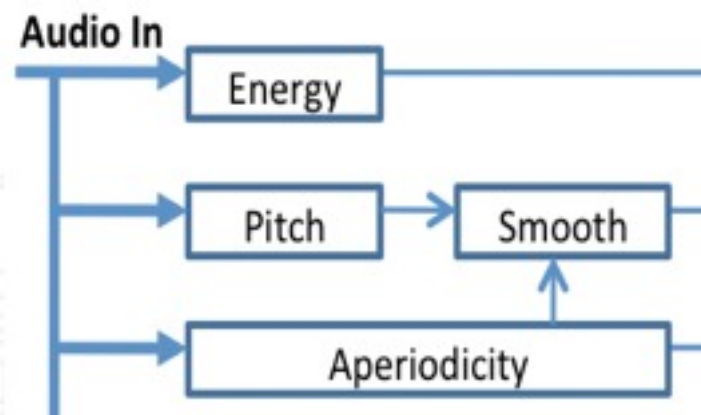
Synthesis

- Sample concatenation
 - **Phase alignment**
 - Offset tests with correlation
- Minor pitch adjustments
 - **Interpolation**
 - Simple
 - No significant artifacts for small pitch changes
 - **Pitch smoothing**
 - Better continuity during overlapping
- Frame energy for dynamics



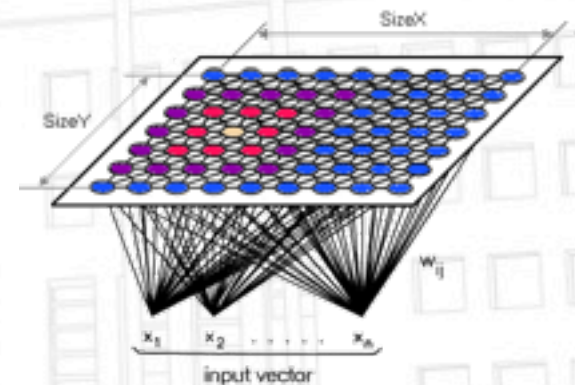
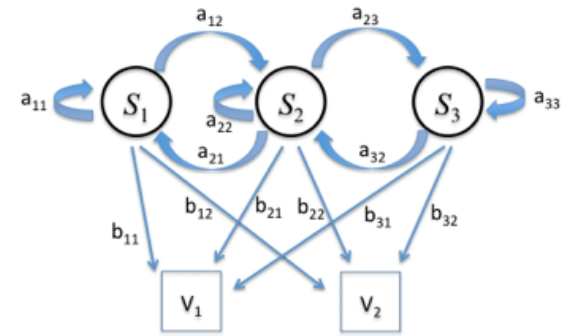
Pitch and Dynamics Extraction

- Pitch extraction based on YIN method
 - YIN method
 - Median smoothing
 - Aperiodicity evaluation and decision
- Dynamic information based on energy
 - Simple, but efficient



Replicating Phonemes

- Several possibilities tested:
 - Phoneme extraction
 - NN, SVM, HMM
 - Phonetic typewriter
 - NN SOM
 - Phonetic similarity
 - Euclidean distance
- Within “Singing Resynthesis”,
phonetic similarity presented the
best results



$$d = \sqrt{\sum_{i=1}^k (X_i - Y_i)^2}$$



Measuring Phonetic Similarity

- Target and Concatenation Cost
- Target Cost:
 - Euclidean Distance with normalized differences within four domains:
 - MFCC
 - LPC Frequency Response
 - LPC Itaruka-Saito Distance (symmetrical)
 - Aperiodicity (YIN)
- Concatenation Cost
 - LPC Frequency Response

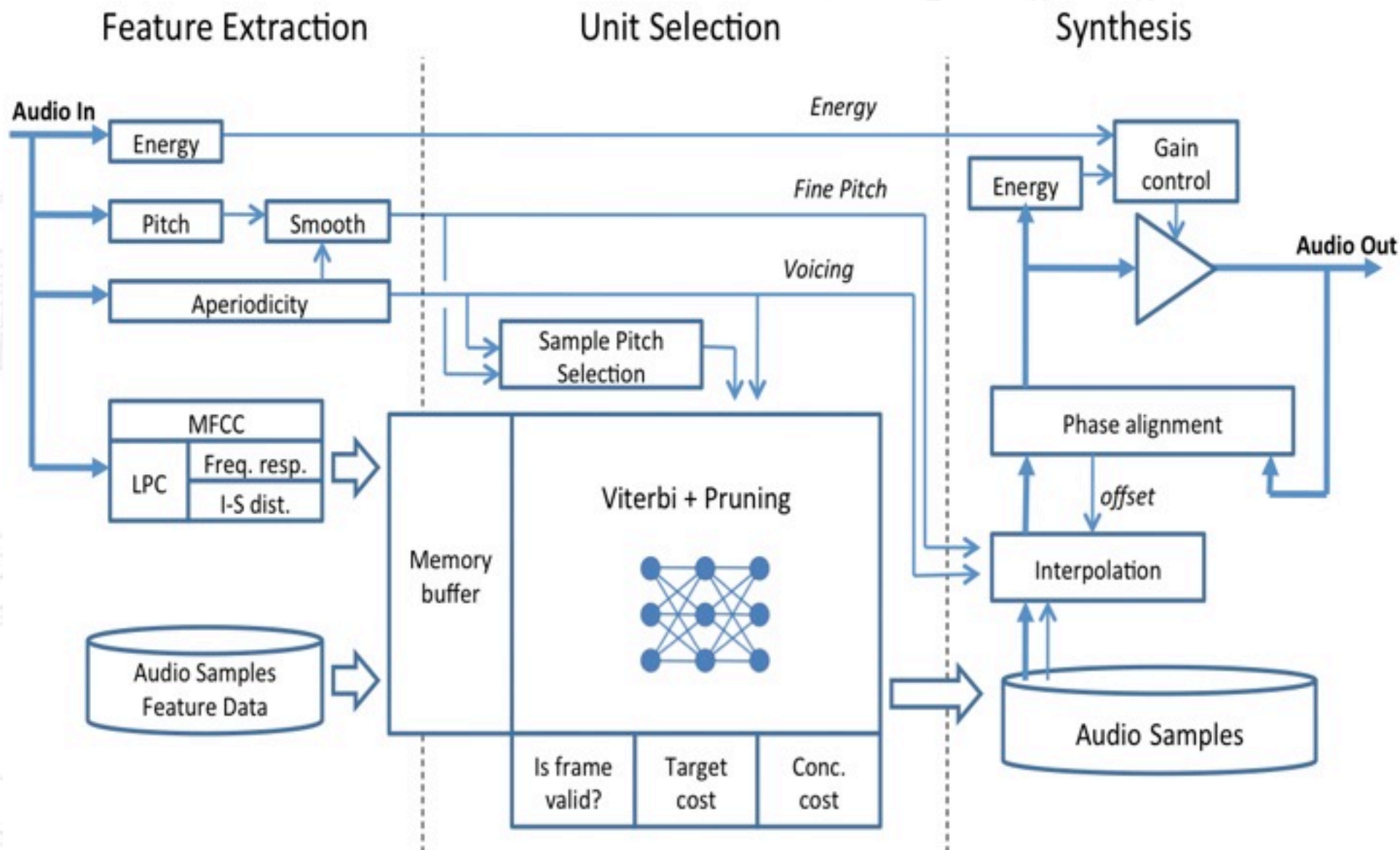


Unit Selection

- Searching for the sequence with the lowest total cost
 - Simple for target cost
 - Complex for both target and concatenation cost
- Several tests, including:
 - Heuristics
 - Heuristics with Segmentation
 - Viterbi
 - Viterbi with pruning
- Final Solution: Viterbi with 10% pruning

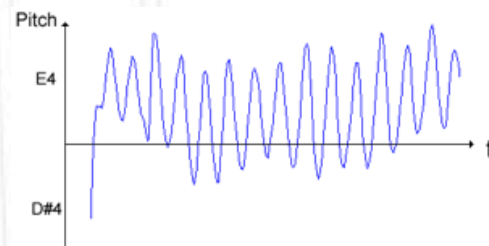


Proposed Method



Improving Sound Quality

- Reducing number of transitions due to pitch
 - Transitions forced by pitch variations
 - Pitch tolerance: 0.5 $\rightarrow$ 1.5 semitone
- Reducing number of transitions due to phonetics
 - 50% more weight on concatenation cost
 - adjacent concatenation cost: 0 $\rightarrow$ - 0.5
- Discarding internal low energy frames
- Prevent frame repetitions
- Considering the effects of time shifts
 - Interpolation/phase alignment will create time shifts $\rightarrow$ incorrect concatenation costs
 - Interpolate worst concatenation scenario



Some examples

- Amazing Grace (LeAnn Rimes)
- Frozen (Madonna)
- Tom's Diner (Susanne Vega)
- Whenever (Shakira)



Conclusions

- Although the concept of Resynthesis is simple
 - Its implementation is very complex
 - High potential for future applications
- Much more research work is required
 - Artifacts still prevent its use on professional applications
- Main obstacles:
 - Inexistence of a singing dataset for research purposes
 - Lack of a numeric metric to evaluate resynthesis results

