Speech Processing 11-492/18-492

Speech Synthesis

Building Voices

Festival Speech Synthesis System

http://festvox.org/festival
General system for multi-lingual TTS
C/C++ code with Scheme scripting language
General replaceable modules
    lexicons, LTS, duration, intonation, phrasing,
    POS tagging tokenizing, diphone/unit selection
General Tools
    intonation analysis (F0, Tilt), signal processing
    CART building, n-grams, SCFG, WFST, OLS
No fixed theories
New languages without new C++ code
Multiplatform (Unix, Windows, OSX)
Full sources in distribution
Free Software

CMU FestVox Project

http://festvox.org

"I want it to speak like me!"

-Festival is an engine, how do you make voices

- Building Synthetic Voices
  - Tools, scripts, documentation
  - Discussion and examples for building voices
  - Example voice databases
  - Step by Step walkthroughs of processes

-Support for English and other languages

-Support for different waveform techniques:
  - diphone, unit selection, limit domain, HMM

- Other support: lexicon, prosody, text analysers

The CMU Flite project

http://cmuflite.org

"But I want it to run on my phone!"

- FLITE a fast, small, portable run-time synthesizer
- C based (no loaded files)
- Basic FestVox voices compiled into C/data
- Thread safe
- Suitable for embedded devices
  - Ipaq, Linux, WinCE, PalmOS, Symbian
- Scalable:
  - quality/size/speed trade offs
  - frequency based lexicon pruning
- Sizes:
  - 2.4Meg footprint (code+data+runtime RAM)
  - < 0.025 secs "time-to-speak"

Festival Speech Synthesis

- *After Installation*
- *festival –tts stuff.txt*
- *festival*
- *festival> (SayText “hello world”)*

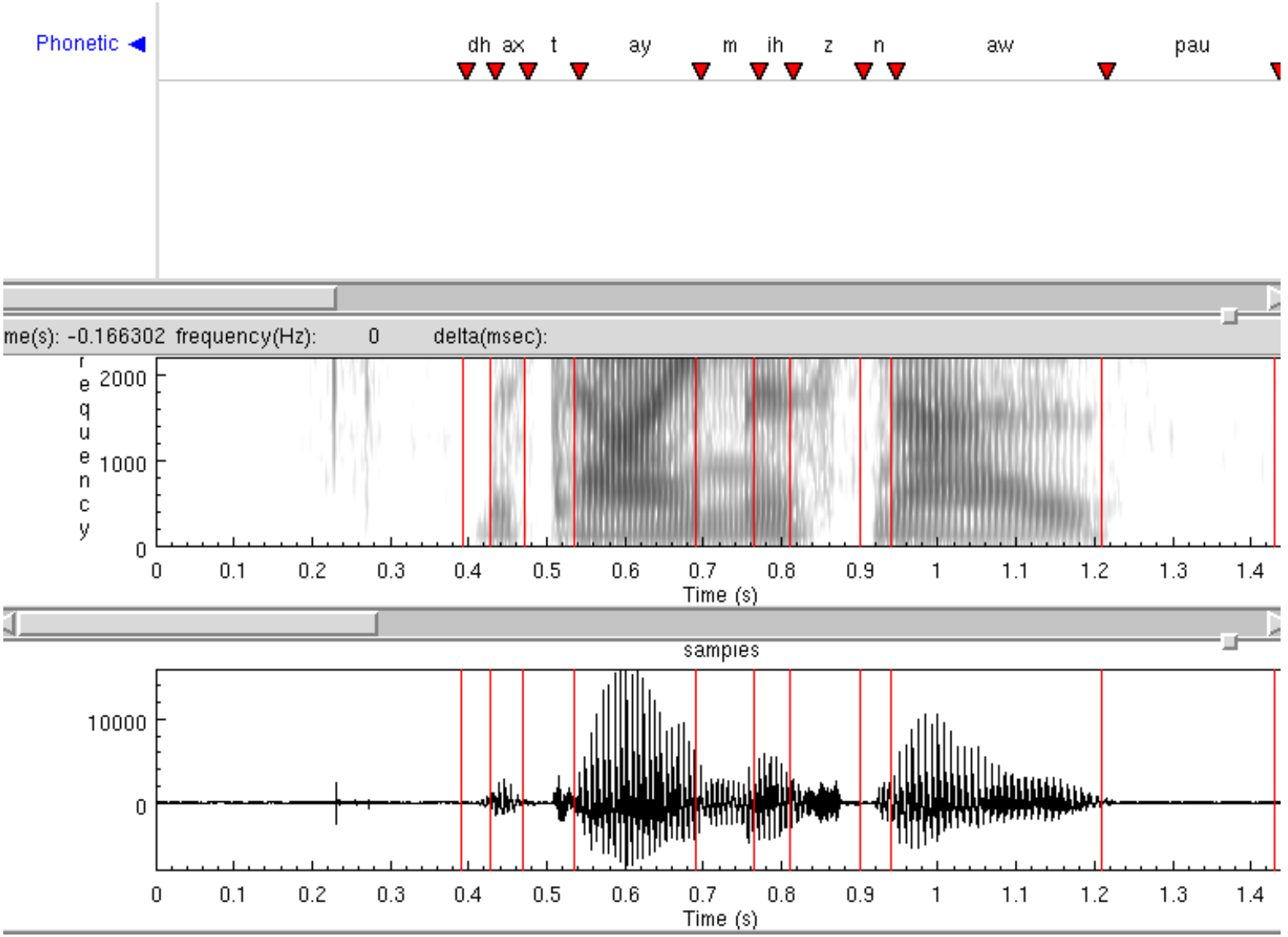
Building Synthetic Voices

- *http://festvox.org/bsv*
  - *Look at section on "Telling the Time"*

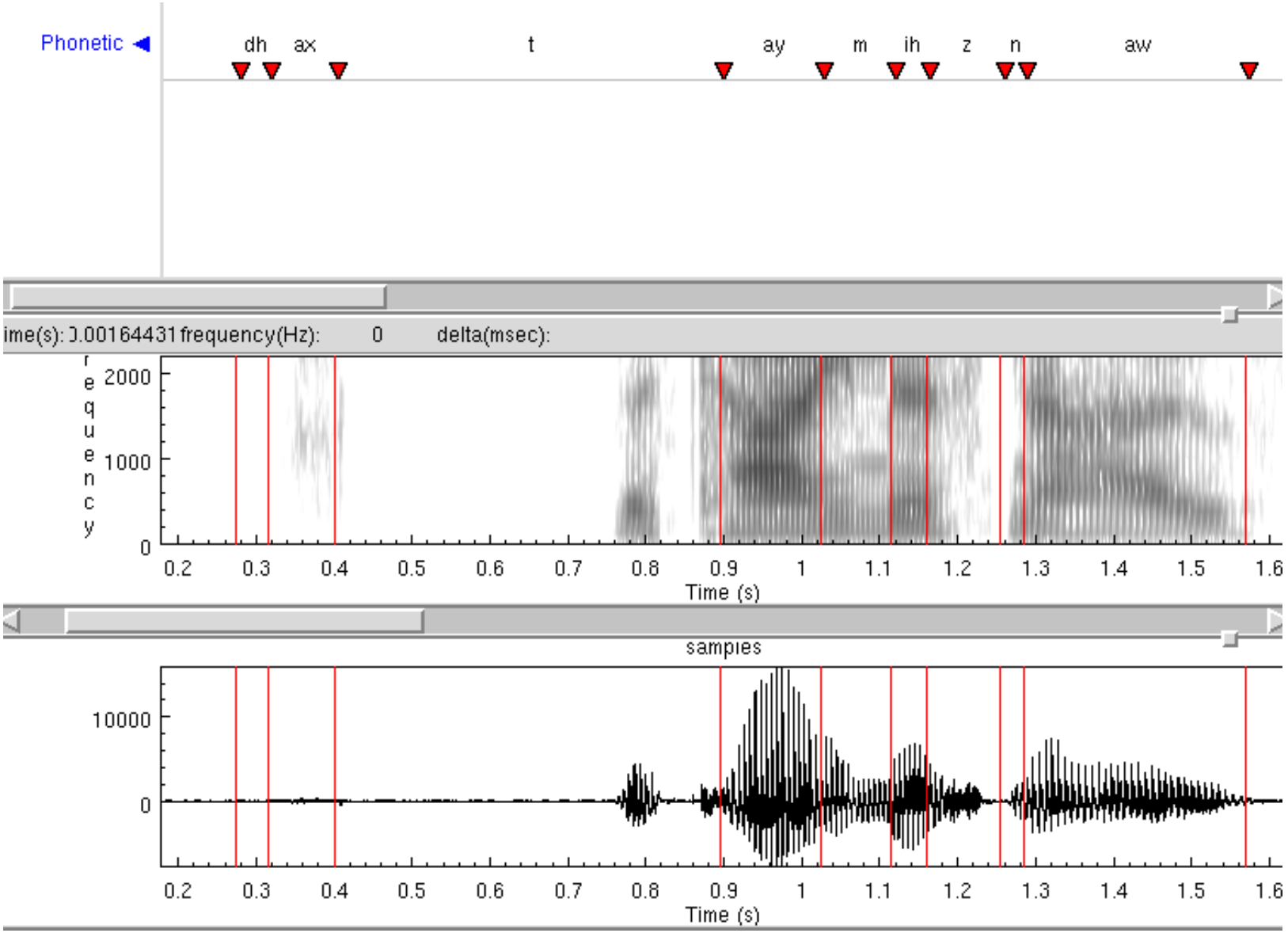
Building a Voice

- *Designing the Prompts*
- *Recording the Prompts*
- *Labeling the Utterances*
- *Finding parameters (F0, MCEP)*
- *Building the synthesis voice*
- *Tuning and Testing*

Automatic Labeling



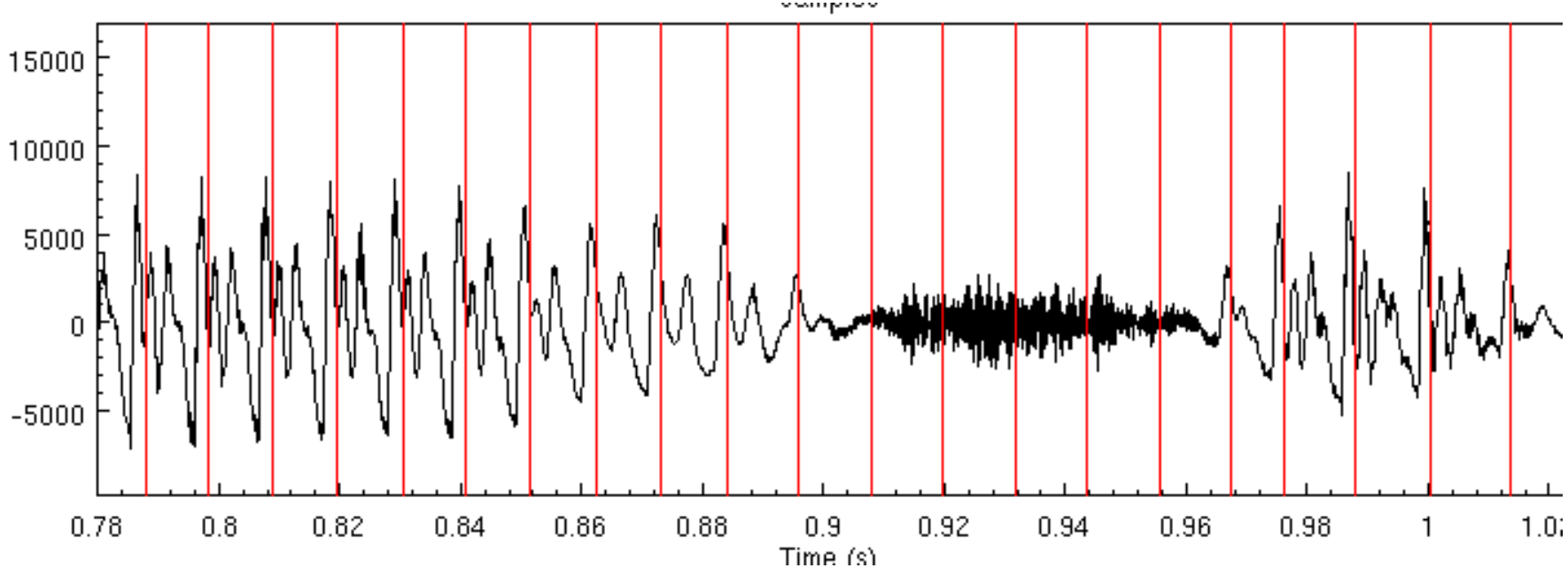
Automatic Labeling (bad)



Parameterization

- *Extract pitch marks from data*
  - *Find voices/unvoiced regions*
  - *Add “fake” pitch marks during unvoiced regions*
- *Extract MFCC pitch synchronously*
  - *Instead of a fixed frame advance (e.g. 5ms)*
  - *Extract it at each pitch mark*
  - *Try to capture the spectrum at the pitch period*

Pitchmarks


15000
10000
5000
0
-5000
0.78
0.8
0.82
0.84
0.86
0.88
0.9
0.92
0.94
0.96
0.98
1
Time (s)

Building a LDOM synthesizer

- *Build cluster tree on each unit type*
  - *Not just on phones*
  - *Tag phones with word they come from*
  - *d_limited and d_domain are treated as different*

Tuning and Testing

- *Test it on some real data*
  - *Ensure number/symbol expansions are correct*
- *Prompts should probably be word expanded*
  - *Flight US187 -> flight u s one eight seven*
- *Remove bad prompts*
  - *Or fix labels*
- *Remember to keep access to the speaker*
  - *If you have to update the system, you need the same speaker available*

Summary

- *Building a voice*
  - *Databases design, recording, labeling*
  - *Parameter extraction and model building*
- *Limited domain synthesis*