Waveform Generation

From phones, durations, F0 to waveforms

Types of synthesis

- Articulartory: model the human vocal tract
- Formant: model the voice signal
- Concatenative: diphones, unit selection
- Canned speech

Waveform generation

- Random word/phrase concatenation
- Phone concatenation
- Diphone concatenation
- Sub-word unit selection
- Cluster based unit selection

Concatenative synthesis

- ☐ Select appropriate speech unit
- ☐ Impose desired prosody
- ☐ Reconstruct signal from modifed parts

Quality is usually good, but less flexible than formant or articulatory.

Diphone synthesis

- □ mid-phone is more stable than edge
- □ Need phone^2 number of units:
  - – some combinations don't exist (hopefully)
  - – may include stress, consonant clusters
  - – lots of phonetic knowledge in design
- □ Database relatively small (by today's standards)
  - – around 8 meg for English (16KHz 16bit)

Designing a diphone inventory

Nonsense words

- ☐ Build set of carrier words:
  - pau t aa b aa b aa pau
  - pau t aa m aa m aa pau
  - pau t aa m iy m aa pau
  - pau t aa m ih m aa pau
- ☐ Advantages:
  - easy to get all diphones
  - will be pronounced consistently
  - (no lexical interferance)
- ☐ Disadvantages:
  - (possibly) bigger db
  - will be pronounced consistently
  - (speaker becomes bored)

As we will be randomly joining these units consistency is probably key

Designing a diphone inventory

Natural words

- ☐ Greedily select sentences/words:
  - quebecois arguments (19)
  - brouhaha abstractions (18)
  - arkansas arranging (11)
- ☐ Advantages:
  - will be pronounced naturally
  - easier for speaker to pronounce
  - smaller db ? (505 pairs vs 1345 words)
- ☐ Disadvantages:
  - will be pronounced naturally
  - may not be pronounced correctly

Diphone distribution in natural text is very variable

Making recordings consistent

Natural words

- ☐ Diphone should come from mid-word
  – help ensure full articulation
- ☐ Performed consistently
  – constant pitch, power, duration
- ☐ Use (synthesized) prompts:
  – help avoid pronunciation problems
  – keep speaker consistent
  – used for alignment in labelling

Building diphone schema

- Find list of phones in language:
  - plus interesting allophones
  - stress, tones, clusters, onset/coda etc
  - foreign (rare) phones,
- Build carriers for:
  - consonant-vowel, vowel-consonant,
  - vowel-vowel, consonant-consonant,
  - silence-phone, phone-silence,
  - other special cases
- Check the output:
  - list *all* diphones and justify missing ones
  - *every* diphone list has mistakes

Recording conditions

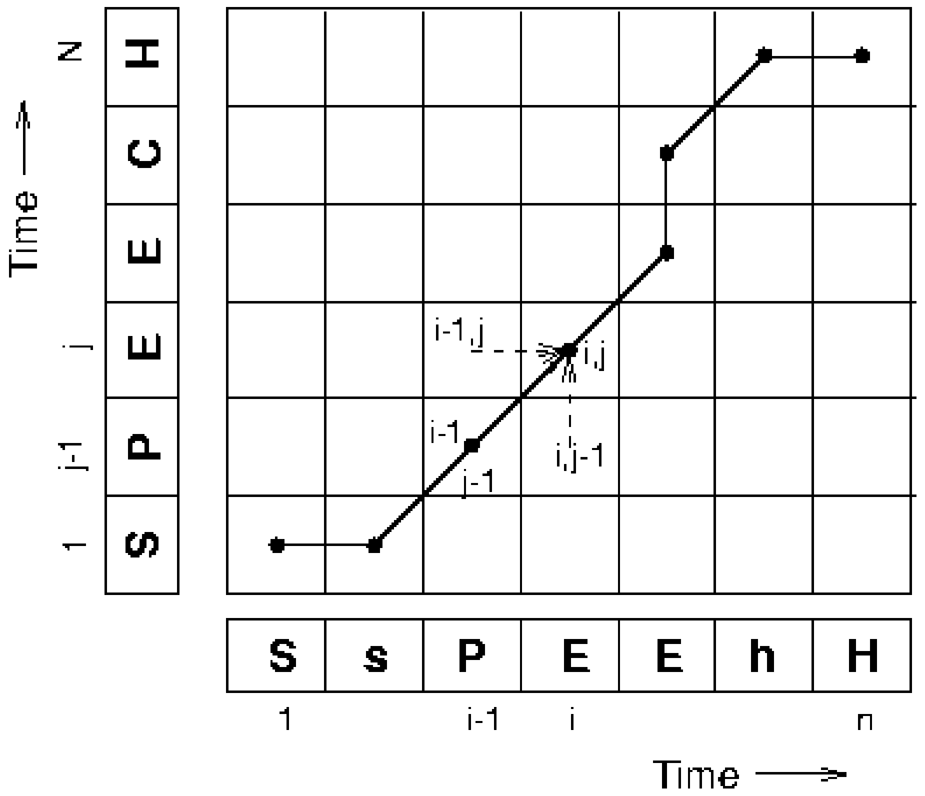
- Ideal:
  - anechoic chamber
  - studio quality recording
  - EGG signal
- What we put up with:
  - quiet room
  - cheap microphone/sound blaster
  - no EGG
  - headmounted microphone
- What we can do
  - repeatable conditions
  - careful setting on audio levels

Labelling Diphones

- *Much* easier than phonetic labelling:
  - the phone sequence is defined
  - they are clearly articulated
  - if its wrong, its wrong
- Phone boundaries less important
  - +/- 10ms is okay.
- Midphone boundaries important
  - where is the stable part
  - can it be automatically found

Dynamic Time Warping

Find shortest euclidean distance through table



Simple autoalignment

Much easier than full autolabelling

- Synthesizer phone string
- Time align prompt to spoken form
  – using euclidean distance
- Works very well 95%+
  – errors are typically large (easy to fix)
  – maybe even automatically detected
- This works cross-language too:
  – even when phones don't exist
  – e.g. English prompts with Korean spoken form

Malfrere and Dutoit 97

Diphone alignment

Does it work?

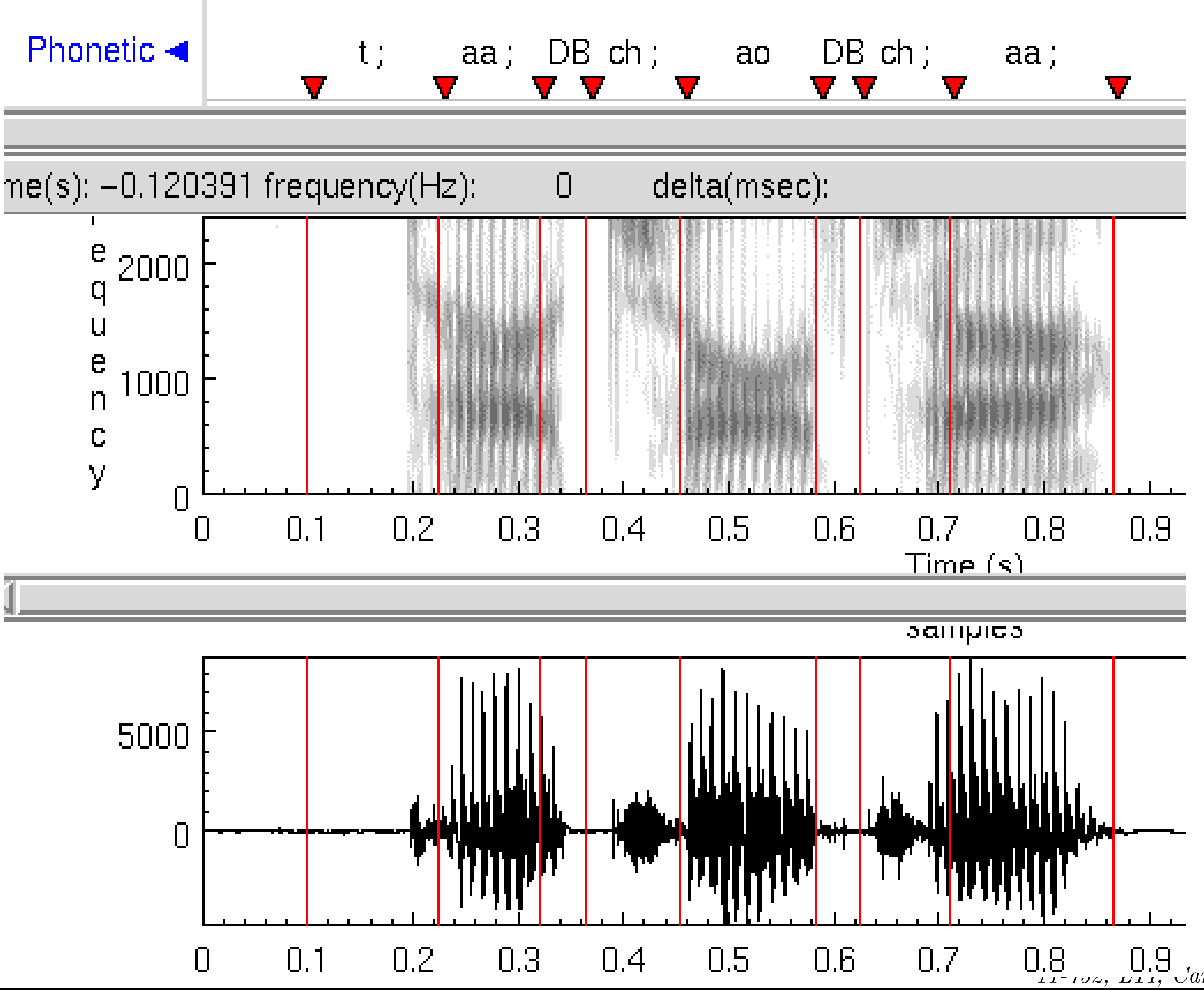
- DP align MFCC prompt to spoken word
- test against hand labelled

	type	RMSE	stddev
KED-KED	self	14.77ms	17.08
MWM-KED	US-US	27.23ms	28.95
GSW-KED	UK-US	25.25ms	23.92
KED-WHY	US-Kor	28.34ms	27.52

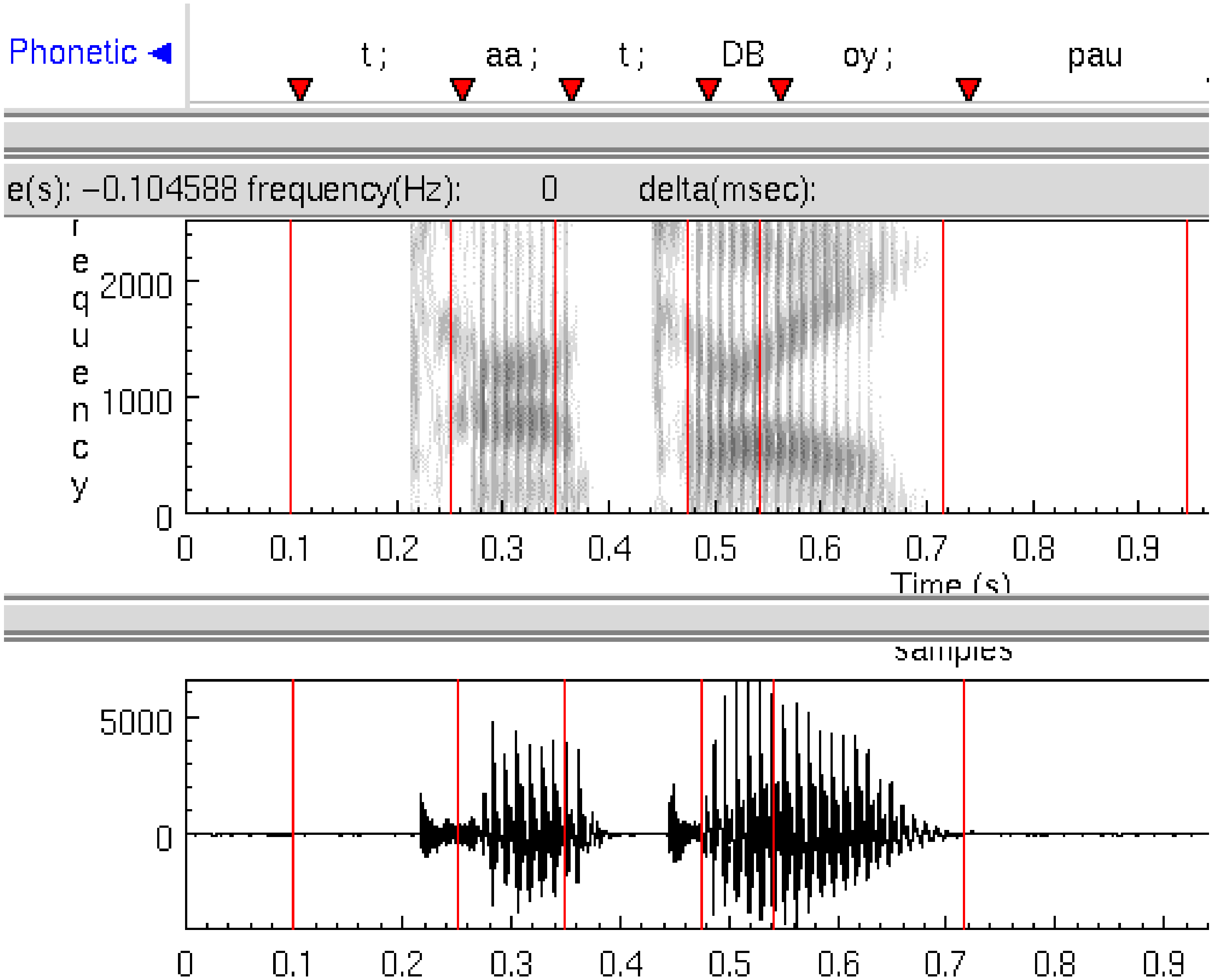
Stable part in phones

- Middle of phone:
  - one third in for stops
  - one quarter in for phone-silence
  - half way for rest
- In time alignment case:
  - Add explicit diphone boundaries
  - (only need to hand correct once)
- Optimal coupling (Conkie and Isard 96)
  - automatically find them
  - using Euclidean distance of cepstrum
  - find minimum join point over all phone-phone
  - or find best for each phone-phone
- Hand check each one:
  - what "real" companies do

Diphone bounadries in stops



Diphone bounadries at end phones



Autolabelling vs Hand labelling

Recorded KAL (US male)

- □ around 15-20 examples *wrong* (KED-KAL)
- □ As good as first pass by human labellers
- □ 45 mins vs 2 weeks hand labelling
- □ Whole voice in under 2 days
  - ◦ recording 3-4 hours
  - ◦ pitch mark extraction 3 hours
  - ◦ alignment 1 hour
  - ◦ hand correction and tuning (3 hours)

But diphone synthesis is too restricted

- Phonetic phonomena go over more than two phones
- Phone-only systems ignore:
  - prosody, stress, syllable position etc
- Two directions:
  - Larger DB
  - More natural DB

Larger database

- ☐ triphones:
  – where it matters
- ☐ stress, onset/coda
- ☐ demi-syllables:
  – approx 10K syls in English

Gives larger, more carefully constructed db:
– more difficult to collect

More natural database

- natural speech has natural coverage:
  - lots of examples of common combines
  - few examples or rare ones
- Should be good for synthesis, if:
  - has basic coverage
  - you can find appropriate units

Why automatic unit selection

- Carefully designed dbs:
  - speaker makes errors
  - speaker doesn't speak intended dialect
  - require db design to be right
- If its automatic:
  - labelled with what was actually said
  - flaps, schwas, coarticulation is natural
- Can better model speaker:
  - want the system to sound like Walter Cronkite
  - picks up ideolect of speaker

Small projects

Presentations week of 5th May

- ☐ New language
- ☐ Limited domain synthesizer
- ☐ New prosodic style
- ☐ Text analysis
- ☐ Translator

May be done in groups

Previous projects

- □ LTS for Thai
- □ Catalan diphone synthesizer
- □ Polish/Chinese weather reports
- □ Fortune teller
- □ Singing Synthesis

Exercises for April 16th

Due noon April 16th

- Build a diphone prompt list for Spanish.

Hints

- Some relevant files in SPPDIR/data/diphones/
- use the given Spanish phoneset definition (though you may consider adding accent vowels too)
- See "Defining a diphone list" at festvox.org for more information `http://festvox.org/bsv/x2245.html`
- You need to generate the list in the format of kaldiph.list
- You may use any programming language to generate it but it *must* be in the right format.