Perfect Synthesis for all of the people all of the time

Alan W Black
Carnegie Mellon University
and Cepstral, LLC

Unit selection synthesis

Concatenate appropriate units from databases of natural speech.

Many dimensions to this problem

□ What data is necessary in the database

□ How much data

□ What should the unit size be

□ What do you do if there isn't an appropriate unit

Unit Selection extremes

□ Diphones:

– one occurrence of each type (phone-phone)
– trivial selection
– requires prosodic modification

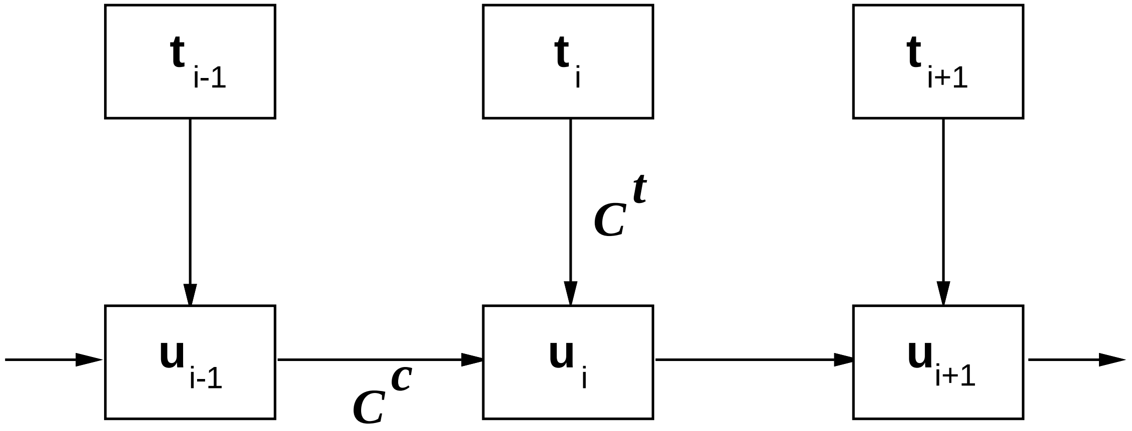
□ "General" unit selection:

– many occurrences of each type
– careful selection (acoustic/phonetic) based
– no prosodic modification

Unit Selection costs

Find candidate units

Find best selection through these candidates



Unit Selection

Target cost: closeness to desired unit
Continuity cost: how well do they join

Find units which minimize:

$$\begin{aligned} C(t_1^n, u_1^n) \quad &= \quad \Sigma_{i=1}^n C^t(t_i, u_i) \;+\; \Sigma_{i=2}^n C^c(u_{i-1}, u_i) \;+ \\ & \quad C^c(S, u_1) + C^c(u_n, S) \end{aligned}$$

“Internal” issues

□ How do we define target costs:
– features and weights

□ How do we score joins:
– acoustic measure matching perception

□ Speech parameterization
– perceptually correlated dimensions

□ Selection algorithms:
– how can you compare them

□ How can do this efficiently:
– clusters, pre-indexing etc

Lots of work to be done here

“External” issues

Find units which minimize:

$$\begin{aligned} C(t_1^n, u_1^n) \quad &= \quad \Sigma_{i=1}^n C^t(t_i, u_i) \ + \ \Sigma_{i=2}^n C^c(u_{i-1}, u_i) \ + \\ &\quad C^c(S, u_1) + C^c(u_n, S) \end{aligned}$$

How can you satisfy this equation well?

Get enough data

□ Record more data:
– cover all possible conditions

But ...

□ Combinatorics are huge:
– “Rare events are common”

□ Humans can’t speak for ever:
– varies over time, not consistent

The right data

Only collect the data you need

□ Find out which data is acoustically different
  – find distances between different unit instances

□ Find out the how often they are needed:
  – looking at very large corpora

□ Find minimal sets that cover the space

But ...

□ Wont be fully general

□ Speaker might not say what you want

Limit your domain

Only synthesize things you can synthesize

□ Can be very high quality

□ Design your database to cover domain:
  - can be infinite domain
  - but constrained, phonetically and prosodically

But ...

□ Is domain specific:
  - maybe ok for your applications

□ Must be easy to build or not worth it:
  - not useful if takes 5 man-years to build

General voice vs Weather voice

Synthesizing in Style

Varying style in the voice:

□ Explicitly record different styles

For examplem database recorded as ...

He _did then _know what _had occurred.
_Tarzan and _Jane raised _their heads.
...

Synthesize as:

This is a short example
_This is a short example
This _is a short example
This is _a short example
...

Change expectations

- □ Make people expect a robotic voice:
  – robots should have robot voices
- □ Make it so it should be hard to understand
- □ Give it a foreign accent

Unit size

□ Word/phrase:

– very large coverage or small domain

□ Phone/diphone:

– easier to get coverage (except for “toy oysters”)

□ Half phone

□ HMM state sized

Boundary positions

□ at *boundary* points

– most dynamic place

– use optimal coupling for mibpoint joins

□ at *stable* points

– cf diphone

Finite vs Infinite number of units

Sounds good if you have the right unit
But if you don’t ...

□ Smooth the joins:
- lightly (power/ pitch period)
- Interpolation (fusion units)

□ Smooth the units:
- HMM generation

Will still be based on the acoustic space of our database

Some of the people all of the time

Some people don’t need high quality

□ “Unnatural” tasks:

- very high speed audio output
- screen readers

□ “Should sound robotic”:

- don’t want natural voice

□ Some people genuinely don’t care

□ Listen often, sounds good

All of the people some of the time

Domain synthesis

□ Design the voices for the tasks:
  - very high quality

□ Limited domains:
  - weather, dialog systems etc

□ Domain directed:
  - say anything but good at most common expressions

□ Style directed:
  - appropriate voice quality
  - command vs compassionate

All of the people all of the time

Far from achieving this

□ Not just good sounding but *appropriate*

– appropriate prosody/style

– not confusing

– can’t evaluate in isolation

□ Even fully natural voices can be disliked:

– personal tastes

– can listener control the voice

– “speak up a bit”

– “don’t be so happy when my stocks have crashed”

□ How can we ever tell?

– evaluation still one of the hardest problems

There is no single voice that can achieve this

Conclusions

Unit selection works well when

□ we carefully construct it

□ we tune it for the application

To improve it we need to

□ do more work

□ have more control over the speech

□ be able to modify the units