and Cepstral, LLC
Carnegie Mellon University
Alan W. Black

Perfect Synthesizers for all of the people all of the time
Unit selection synthesis

Many dimensions to this problem

What data is necessary in the database?

What should the unit size be?

What is the unit size of an appropriate unit?

How much data?
no prosodic modification
- careful selection (acoustic/phonetic) based
- many occurrences of each type

- General unit selection:
- requires prosodic modification
- trivial selection
- one occurrence of each type (phone-phone)

Unit Selection extremes
Find best selection through these candidates

Find candidate units

Unit Selection costs
Find units which minimize:

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

Unit Selection
Lots of work to be done here

- clusters; pre-indexing etc
- How can we do this efficiently?

- how can you compare them

Selection algorithms:

- perceptually correlated dimensions
- Speech parameterization

- acoustic measure matching perception

- How do we score Jones?

- features and weights

- How do we define target costs?

“Internal” issues
How can you satisfy this equation well?

\[
(S', w)\mathcal{C} + (\Gamma_n', S)\mathcal{C} + \left(\begin{array}{cc} n' & -1 \\ -1 & n' \end{array}\right)\mathcal{C} \Rightarrow \left(\begin{array}{c} \Gamma_n' \\ \Gamma_n' \end{array}\right)\mathcal{C} = \left(\begin{array}{c} \Gamma_n' \\ \Gamma_n' \end{array}\right)\mathcal{C}
\]

Find units which minimize:

„External „ Issues
Get enough data

- varies over time, not consistent
- Humans can't speak for ever: □
- Rare events are common
  - Combinatorics are huge: □
  - But ...
- cover all possible conditions
- Record more data: □
Speaker might not say what you want. □

Wont be fully general. □

... But □

Find minimal sets that cover the space. □

Finding at very large corpora –

Looking at how often they are needed. □

Find out how different unit instances –

Find out which data is acoustically different. □

Only collect the data you need.

The right data.
General voice vs. Weather voice
not useful if takes 5 man-years to build
Must be easy to build or not worth it:
maybe ok for your applications
Is domain specific:

... But

but constrained, phonetically and prosodically

... can be infinite domain

Design your database to cover domain:
Can be very high quality

Only synthesize things you can synthesize

Limit your domain
... This is a short example
This is a short example
This is a short example
This is a short example
Synthesise as:
... Tarzan and Jane raised their heads.
He did they know what had occurred.
... For example, database recorded as

Explicitly record different styles
Varying style in the voice:

Synthesising in Style
Give it a foreign accent

Make it so it should be hard to understand

— robots should have robot voices

Make people expect a robotic voice

Change expectations
ct diphyne

at stable points

use optimal coupling for midpoint joins

most dynamic place

at boundary points

Boundary positions

HMN state sized

Half phone

Phone/diphyne:

easier to get coverage (except for "toy oysters")

Word/phrase:

very large coverage of small domain

Unit size
WIII still be based on the acoustic space of our database

- HMM generation
  - Smooth the units □

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

... But if you don’t

Sounds good if you have the right unit

Finitie vs Infinite number of units
Listen often, sounds good.

Some people genuinely don't care. 

I don't want a natural voice. 

Should sound robotic. 

Screen readers. 

Very high speed audio output. 

"Unnatural" tasks. 

Some people don't need high quality.

Some of the people all of the time.
command vs. compassionate

appropriate voice quality

- Style directed:

say anything but good at most common expressions

- Domain directed:

weather, dialogue systems etc.

- Limited domains:

very high quality

- Design the voices for the tasks:

Domain synthesis

All of the people some of the time
There is no single voice that can achieve this – evaluation still one of the hardest problems

- How can we ever tell?
- "don't be so happy when my stocks have crashed"
- "speak up a bit"
- can listener control the voice
- personal tastes

Even fully natural voices can be disliked:
- can't evaluate in isolation
- not confusing
- inappropriate prosody/style
- Not just good sounding but appropriate

Far from achieving this

All of the people all of the time
Conclusions

To improve it we need to

we carefully construct it

when selection works well

we tune it for the application

we do more work

have more control over the speech

be able to modify the units