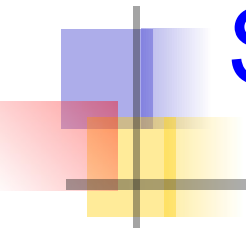


SyRG Review, May 27 2004

Optimal Data Selection for Unit Selection Synthesis, 2001

A. Black, K. Lenzo

Gist of Paper

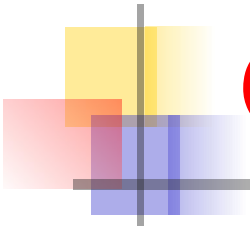
- Goal
  - Find an optimal prompt set to record
    - for unit selection voices
    - designed towards targeted open domains
    - captures acoustic-phonetic range of speaker
- Innovation
  - use acoustic coverage as selection criteria
    - call these “clustered acoustic units”
  - use statistical distribution of cluster units

Optimality

- What is optimal?
  - see last paragraph of Section 2
  - Full Coverage
    - have examples of everything you need
  - Minimal Redundancy
    - without unnecessary recording
  - Working Definition
    - voice gets worse if you remove prompts
    - doesn't get much better if you add prompts
      - e.g. talking clock limited domain

Trouble on the Horizon

- Results
  - proposed method selects 241 prompts
    - hand pruned down to 221
      - “smaller than we expected” ... “in order to get more examples we ran the selection algorithm again”
    - second set of 146 prompts; combined 347
    - Arctic experience says this is too small
  - Evaluation
    - combined set tested better than smaller pair
    - thus method under-represents speaker



Outlook

- Opinion
  - basic idea is solid
  - parameterization isn't right
  - so what's the deal?
- This talk
  - explore method, propose refinements
  - return to topic again later
    - vet results in a later SyRG meeting

General Constraints

- Unit Selection Synthesis
  - capable of high quality (easier modeling)
  - carries with it the style of recordings
  - stay within domain
    - e.g. not attempting to read stories based on newscast speech
  - don't perform unit modification
    - i.e. voice transformation enables a greater range of output with less recorded material

Text-only Prompt Selection

- Limited Domain
  - start with list of utterances (or generator)
  - greedy select on words
    - synthesize with word-sized units
    - for tighter phonetic control, select and synth words marked with preceding word
      - “word joins may be poor” (s2.2)
- Foreign Language
  - fallback if no letter-to-sound rules

Phonetic-Symbol Selection

- Predict acoustics from text
  - from lexicon and lts rules
    - text to phonemes to units
    - many possible units
      - phones, diphones, triphones, syllables, demisyl
    - plus attributes that affect sound
      - lexical stress, phrase posn, pitch, etc.
      - which factors are important is not known
      - exhaustive coverage impossible to collect [vanSanten 1997]

Coverage vs Distribution

- Complete Coverage
  - at least one example of each unit
  - diphone databases are designed to have exactly one of each (s2.6)
- Natural Distribution
  - frequency of selected units same as domain
  - provide more choices for common usage
  - Lenzo algorithm tries to avoid high frequency selection bias (s4)
    - unnecessarily complicated! [jk]

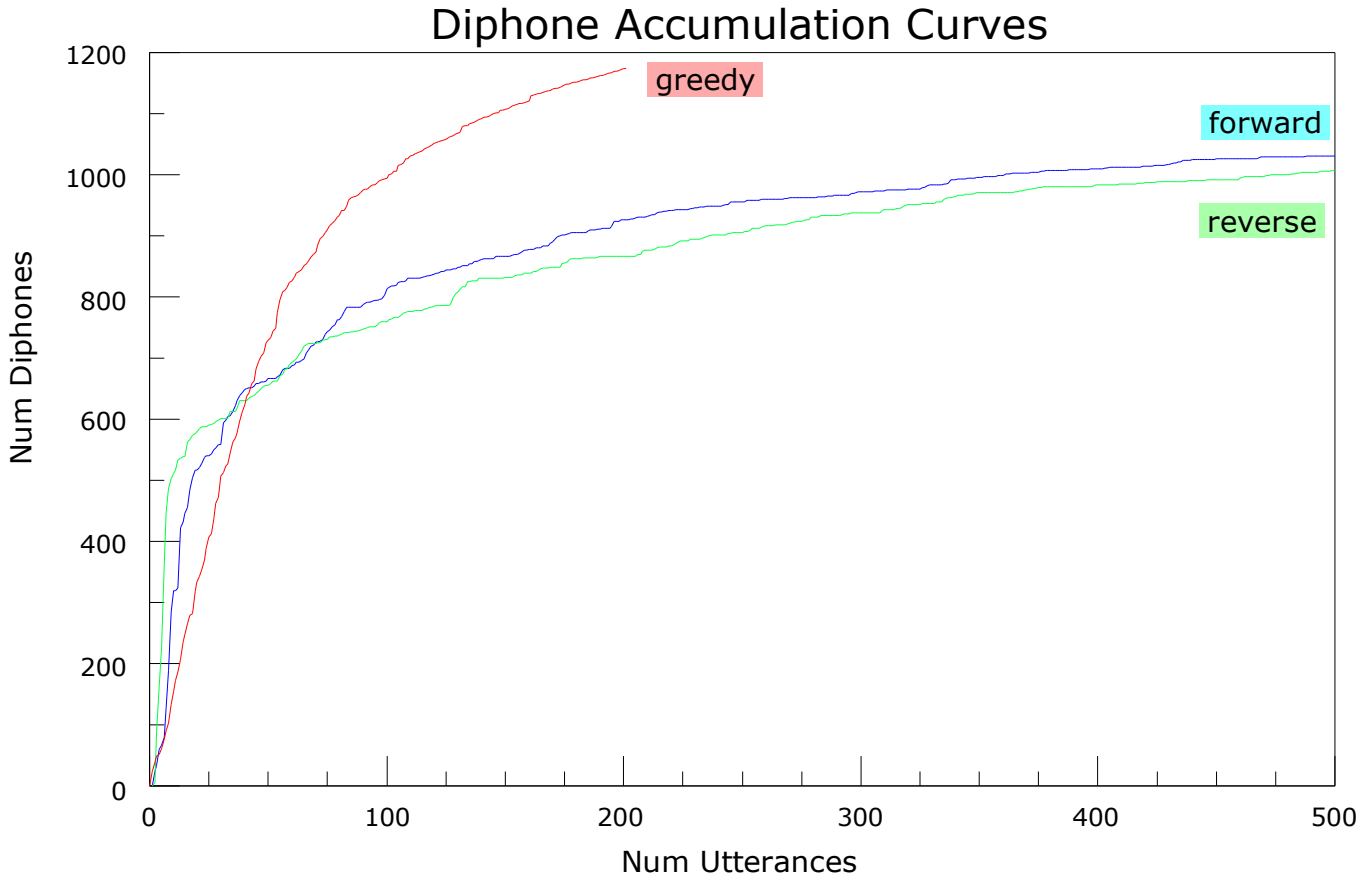
Coverage Volume on AiW

	Batch	Incremental		
	Utts	Utts		Units
Unstressed				
phone	6	3	41	601
diphone	196	192	1174	17405
triphone	1205	1199	10214	74455
Stressed				
phone	10	9	51	1015
diphone	235	229	1366	20376
triphone	1266	1262	10982	76862
words (CS)	894	887	2995	17123
words (CI)	764		2603	15916
di-words	1684		13299	24334
letters		3	27	757
di-letters		80	429	11683
tri-letters		395	3017	46414
	of 1920			

Greedy Algorithm

- Basic idea
  - select items one-by-one that maximally improve the objective measure
  - unit coverage is a packing problem
- Two variants
  - recompute item scores after each selection
  - recompute scores after full insertion sweep
    - second variant is faster
    - second implemented in festvox

Accumulation Curves



Comparing Greedy Variants

- Specification
  - unit type - diphones
  - algorithm
    - 1. iterative at block granularity
    - 2. iterative at utterance granularity
  - utterance scoring
    - num new units
    - (num new units) / (utterance length)

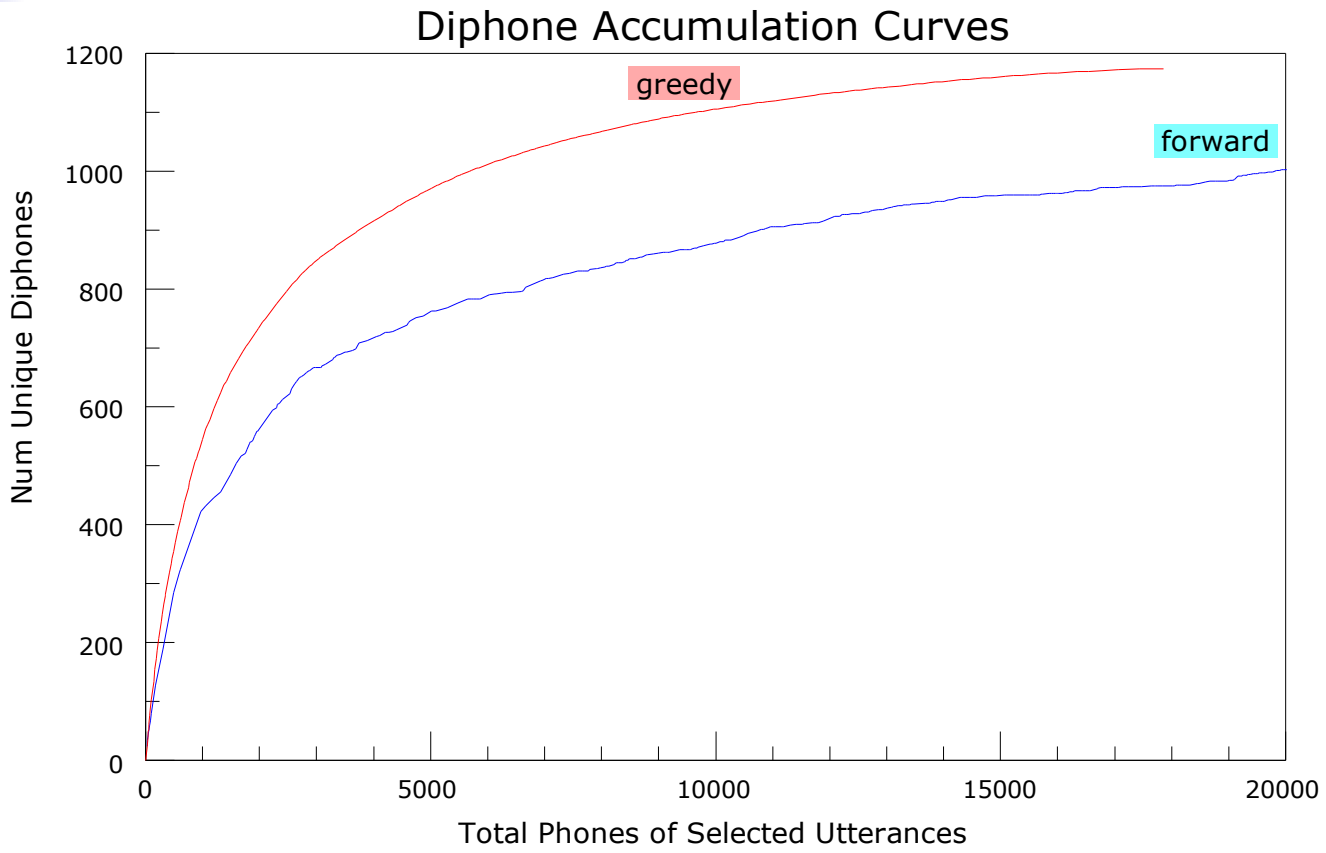
Utterance Scoring

- frequency weighted by units
  - selects common speech (e.g. by diphone count)
- count of new units
  - favors long utterances
- ratio of new units to utterance length
  - favors short utterances
- new unit count x f(utterance length)
  - e.g. hat function H(5,20) used in festvox
  - e.g. Gaussian G(16,8)

Prompt Files Examples

- Compare 3 score functions
  - new item count
  - new item ratio
  - new item ratio with length weighting
- Examples for Alice in Wonderland

Accumulation Curves (2)



How much is enough?

- Problem
  - counting symbols isn't same thing as measuring acoustics
  - relation between two isn't known
  - needed redundancy isn't known
- Black & Lenzo proposition
  - start with augphones as speech units
  - cluster units by phonetic features
  - hypothesis – one example of each is enough

Augmented Phones

- Important detail
  - clustered segments are “augphones”
    - phoneme plus 50% of previous phone
    - Why? Join continuity
    - see ref [3]

Cluster Trees

- Example
  - ((R:SylStructure.parent.syl_break is 4)
    ((n.name is pau)
    ((name is s)
    ((p.ph_cvox is 0)
    ((45 986 324 892))))))
  - If the current phone /s/ is followed by a pause and we are at a large phrase break (val 4), and the previous phone – a consonant – has unknown voicing, then in this context an /s/ is represented by the unit set with id numbers {45,986,324,892}

Distance Metric

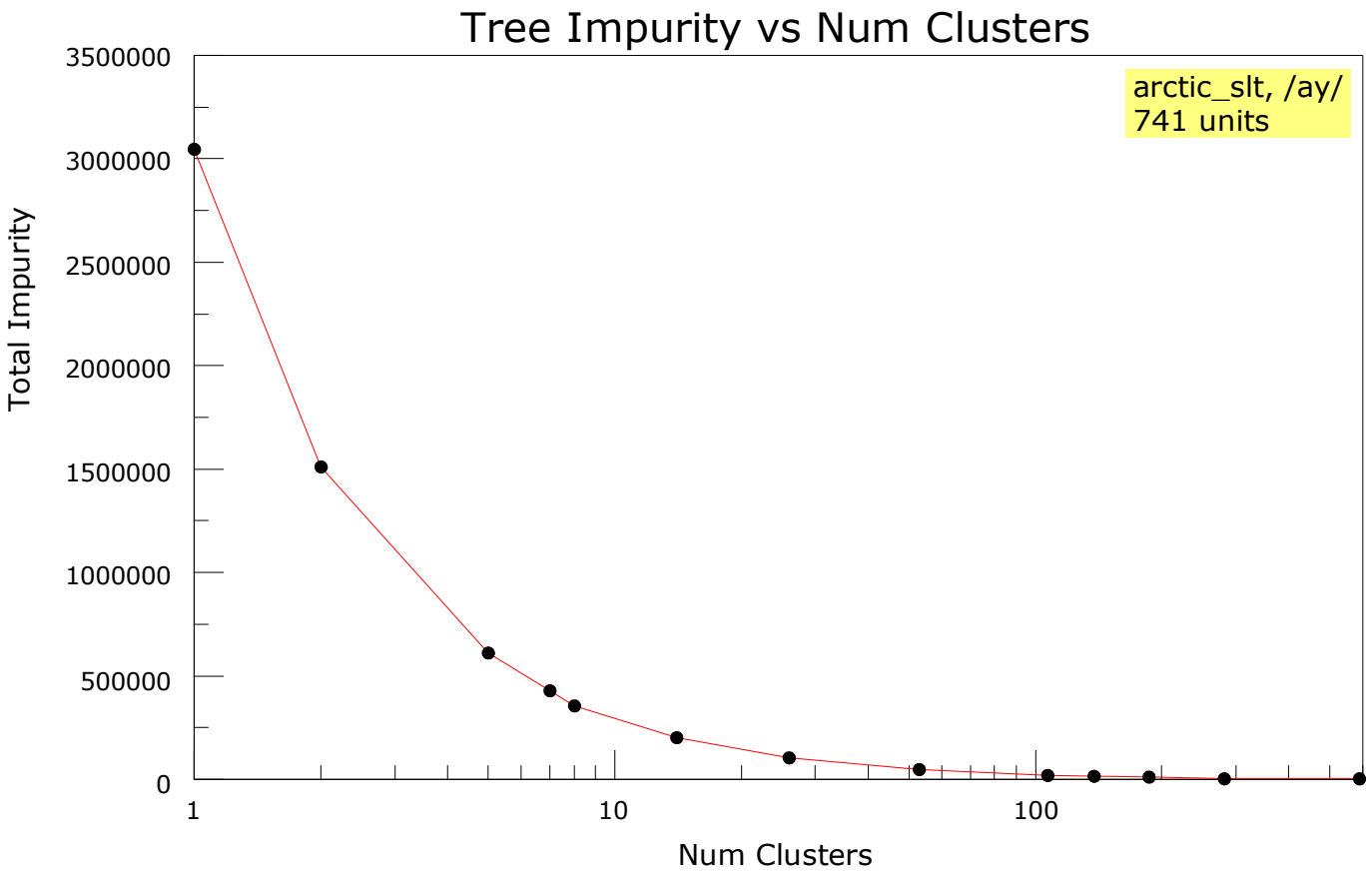
- Distance measure for clustering
  - weighted ceptral frames with length alignment
    - j: iterates n mel frequency cepstral coefficients
    - i: iterates over frames in U
    - σ: stdev for Mahalanobis distance
    - W: weights on ceptral components
    - P: penalty term for length disparity

$$D(U,V)=P(\frac{|U|}{|V|})\frac{1}{n|U|}\sum_{i=1}^{|U|}\sum_{j=1}^{n}\frac{W_j}{\sigma_j}\left|F_{ij}(U)-F_{(i\,round(\frac{|V|}{|U|}))\,j}(V)\right|$$

Impurity vs Cluster Count

- What is the right stopping threshold?
  - balance between cluster purity and number of cluster representatives
  - one example per cluster is too few
  - redundancy needs greater when database has not been hand recorded
- Note
  - selection tree doesn't have to be synthesis tree

Impurity Curve for AY



Alternative Clusterer

- HMM acoustic training
  - use senomes as clusters
    - each tied triphone state represents a distinct phonetic segment
    - problems
      - subphone segments of speech
      - num senomes is a free parameter

Evaluation

- example wavefiles
  - see www.festvox.org/dataselect