Synthesizer Voice Quality of New Languages Calibrated with Mean Mel Cepstral Distortion

John Kominek Language Technologies Institute
Tanja Schultz Carnegie Mellon University
Alan W Black Pittsburgh PA USA

SLTU Hanoi Vietnam – May 6 2008

SPICE project goals

- Rapid development of ASR + TTS for new languages
- SPICE – Speech Processing Interactive Creation and Evaluation
  - a web-based tool suite that streamlines creation of core language technology components
  - Janus multi-lingual ASR
  - Festival multi-lingual TTS
  - text and speech collection
  - phoneme and character set definitions
  - LTS rule builder

CMU SPICE

Build Your System

- Text and prompt selection (help)
- Audio collection (help)
- Phoneme selection (help)
- Grapheme-to-phoneme rules (help)
- Lexicon pronunciation creation (help)
- Build acoustic model (help)
- Build language model (help)
- Test ASR system
- Create speech synthesis voice

User: **john** Language: **eng** Project: **recipe_1000** [Logout]

Building synthesis voice

Tasks

Voice Name: cmu_spice_eng_recipe_1000
Voice Directory: cmu_spice_eng_recipe_1000
Tasks:

- recreate voice (and delete current one)
  cmu_spice_eng_recipe_1000
- import_waves waves/
- import_prompts txt.done.data
- import_lexicon lexicon lexrules
- label_segments lab/
- extract_params ccoefs/
- build_models trees/
- build_dur festvox/
- test_voice
- package_voice festvox_cmu_spice_eng_recipe_1000_cg.tar.gz

Initial evaluations

- Conducted 2 semester-long lab courses
  - students use SPICE to create working ASR and TTS in a language of their choice
  - bonus for the ambitious
    - train statistical MT system between two languages to create a speech-to-speech translation system
- Evaluation includes
  - user feedback on difficulties
  - time to complete
  - ASR word error rate
  - TTS voice quality (this paper)

Focus on TTS

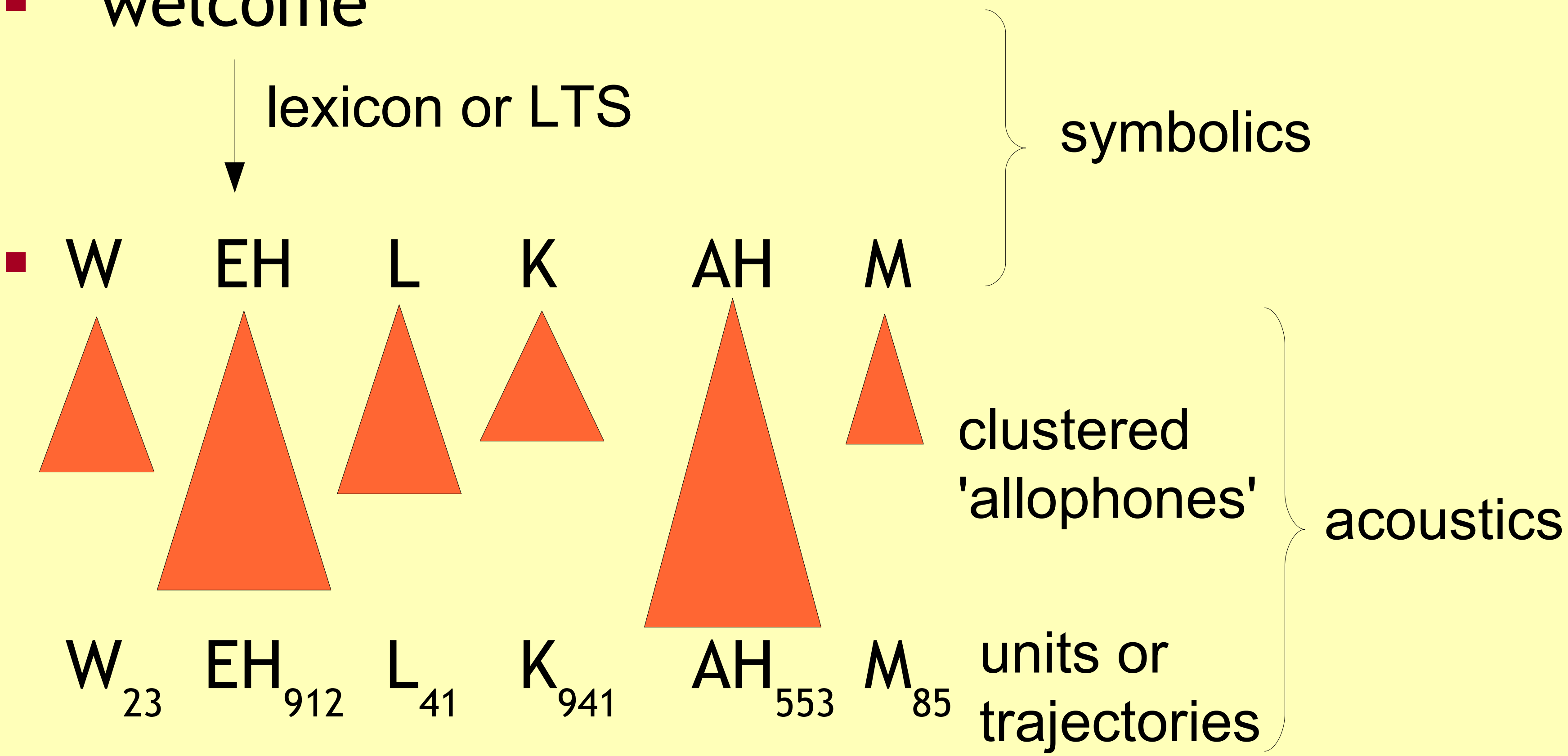
- Main research questions
  ١. To what extent is language-dependent expertise required of the user?
  ٢. To improve the synthesizer, what is the most efficient use of the user's time?
  ٣. How can we measure the user's progress?

Research question in detail

- Language dependence
  ١.Which features matter the most in CART tree training? Are language-dependent features critical?
  ٢. What is the best 'stop value' for training?
- Measurement
  ١. Can an objective measure be used to estimate the quality of a voice, in any language?
  ٢. Can this information motivate and inform the user?
- Efficiency
  ١. Rate of improvement as more speech is recorded?
  ٢. Rate of improvement as the lexicon is expanded and corrected?

TTS overview

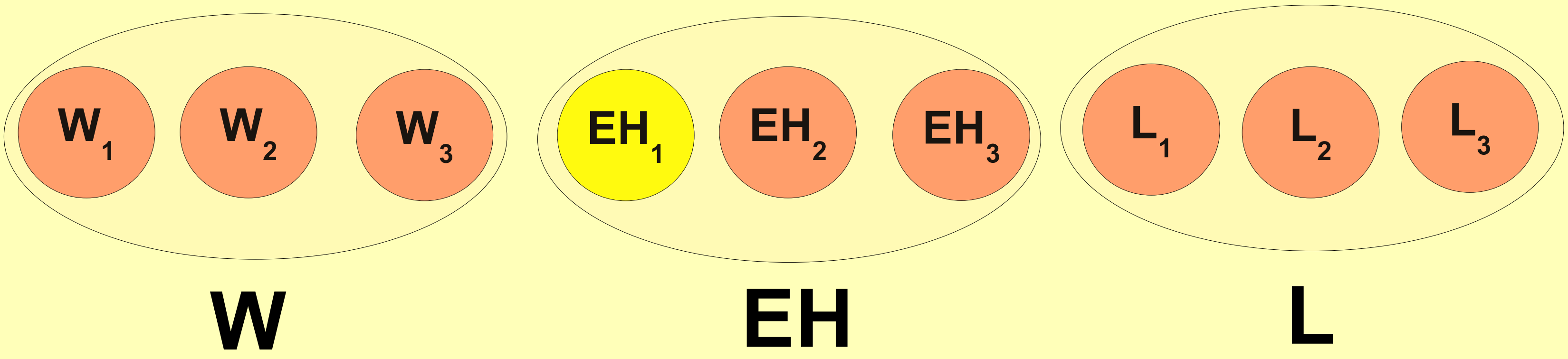
- “welcome”



Key point - quality of CART trees depends on:
training features, amount of speech, label accuracy

Context-dependent CART training

- Suppose text is “hi welcome to”
  - when training the EH_1 state we use name feats
  - prev states: ... AY_3 W_1 W_2 W_3
  - next states: EH_2 EH_3 L_1 L_2 ...
  - prev phones: # HH AY W
  - next phones: L K AH M



More CART tree features

- Four categories of training features
  ١. names: phoneme and HMM state context
  ٢. position: e.g. number of frames from beginning of state, percentage in from beginning
  ٣. IPA: International Phonetic Association features, based on phoneme set
  ٤. linguistic: e.g. parts of speech, syllable structure

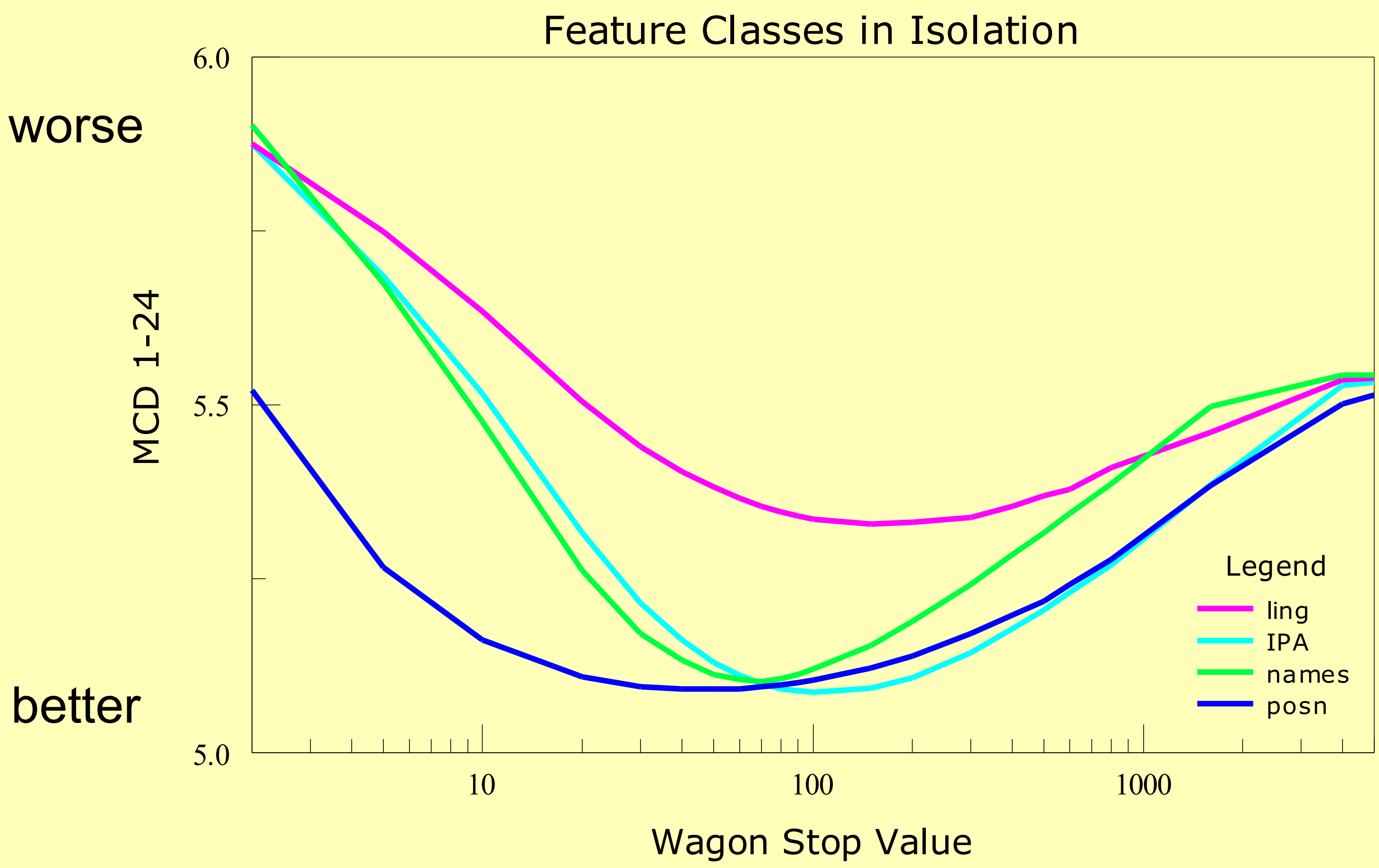
- level of language expertise required
  - 1. and 2. are language-independent
  - 3. requires an IPA-based phoneset
  - 4. requires a computational linguist

Calibration experiments in English

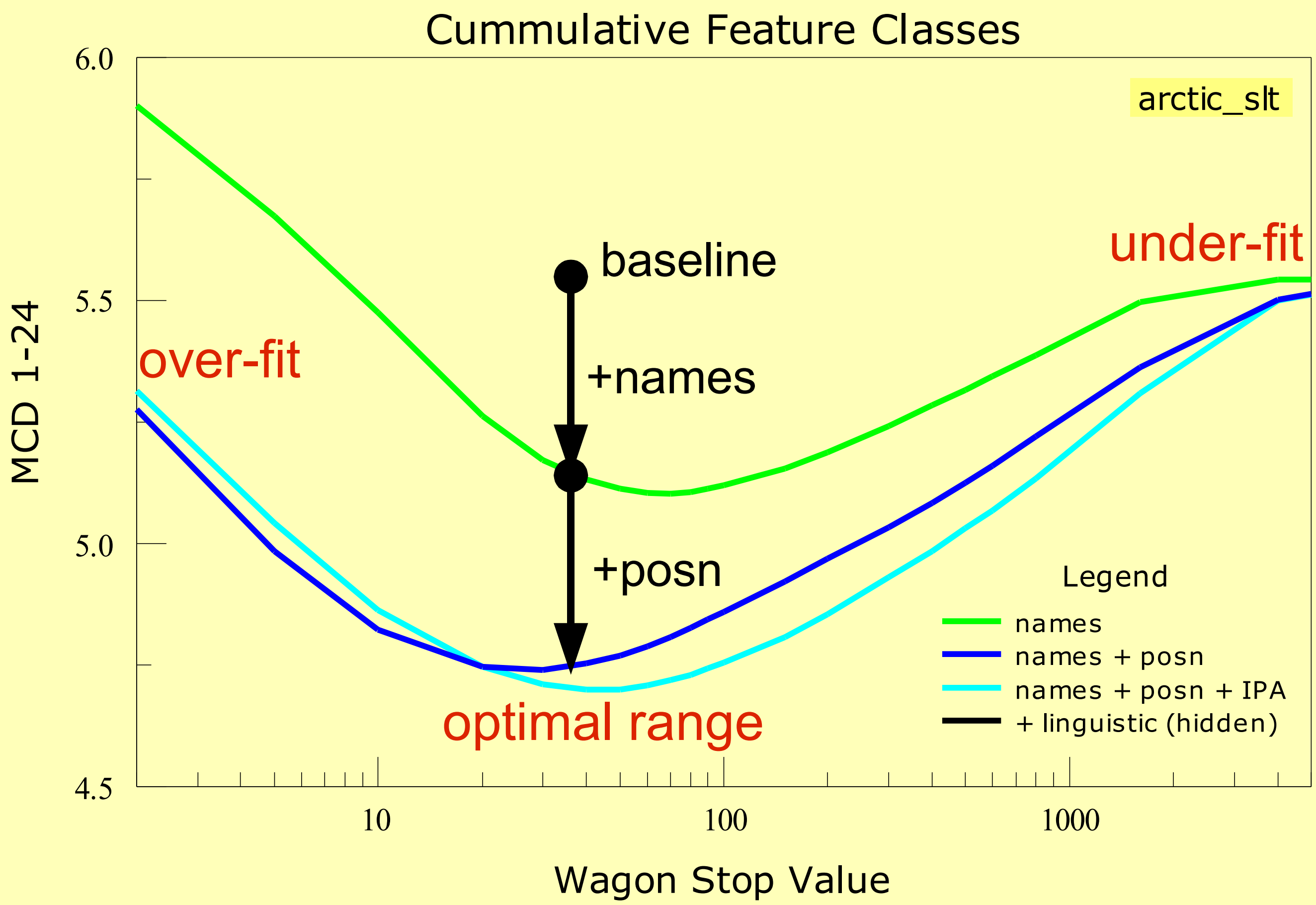
- Use a studio-recorded database (arctic_slt)
  - 1 hour of clean speech
    - 90% training / 10% test – partitioned 10 times into separate testing sets
    - vary the amount of speech used to train
    - vary the CART training features
    - vary the CART stop value
- Compute mean mel cepstral distortion (MCD)
  - average frame-to-frame Euclidean distance between synthesized and original wavefile
  - let v = sequence of 25-D cepstral frames, 5 ms step

$$MCD(v^{targ}, v^{ref}) = \frac{\alpha}{T'} \sum_{\substack{t=0 \\ ph(t) \notin SIL}}^{T-1} \sqrt{\sum_{d=1}^{D} (v_d^{targ}(t) - v_d^{ref}(t))^2}$$

Effect of isolated feature classes



Effect of combined feature classes



Effect of feature classes

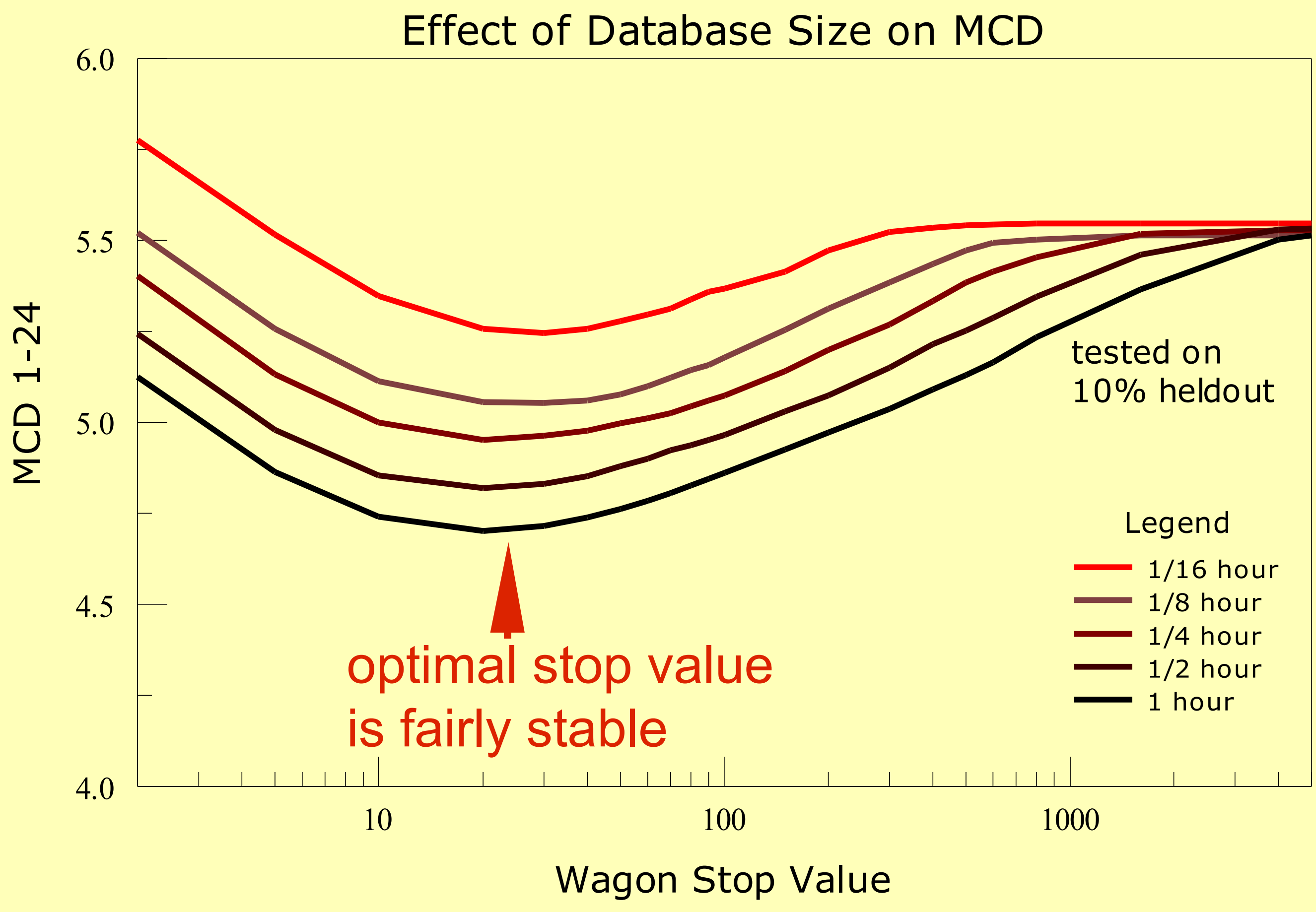
- lower numbers are better
  - ~ 0.2 is perceptually noticeable
  - ~ 0.08 is statistically significant
- the first two feature classes matter
  - from the minimum values of each feature class...

Feature class	***Features***	***Lang dep.***	***Δ MCD***
no CART trees	0	no	baseline
name symbolics	16	no	- 0.452
position values	7	no	- 0.402
IPA symbolics	72	yes	- 0.001
linguistic sym.	14	yes	+ 0.004

Effect of database size

- Doubling speech reduces MCD by 0.12 ± 0.02
  - a consistent result over many data points
  - thus 4x the speech is needed for a definite perceptual improvement
    - i.e. play two voices side-by-side and the larger voice is clearly better
- Exception at small end
  - from 3.75->7.5 minutes MCD drops by 0.2
  - 10 min of speech can be considered the bare-minimum starting point

Effect of database size on MCD curves



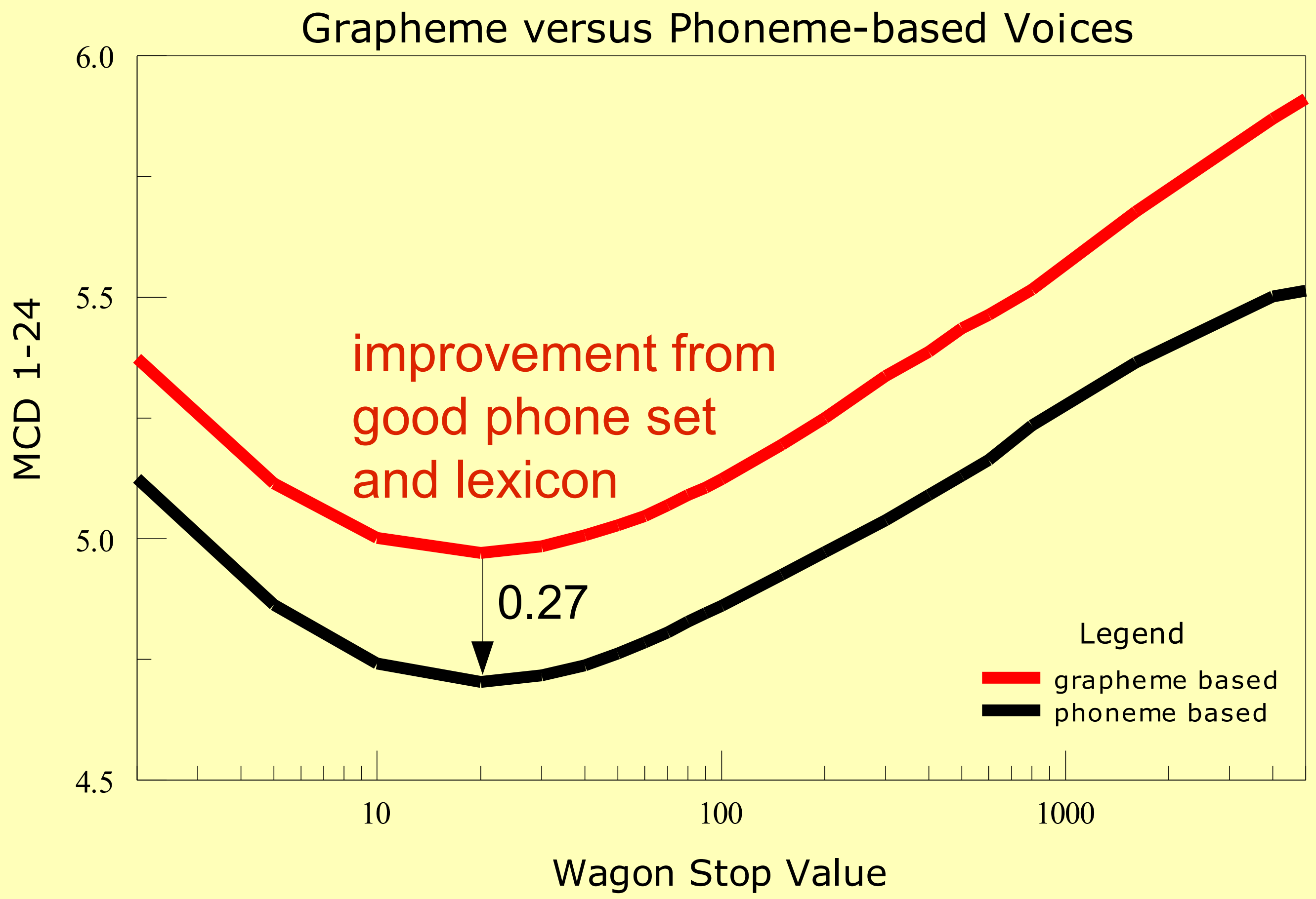
Plenty of room at the high end

- Point of diminishing returns not evident in these experiments
- Where is the asymptote?
  - don't know yet
  - maybe 20 hours of consistently recorded speech
  - however, large databases recorded over multiple days are plagued by inconsistent recordings

Effect of a good Lexicon

- Want to simulate what you get with a sub-optimal phone set and a poor lexicon
- Idea: use a grapheme-based voice
  - 26 letters a-z are a substitute 'phone' set
  - no IPA and linguistics features
  - English has highly irregular spelling
    - the acoustic classes are impure
    - caveat: measuring global voice quality not mispronounced words
- Results
  - MCD improves by 0.27
  - consistent across CART stop value

Grapheme vs Phoneme English voices



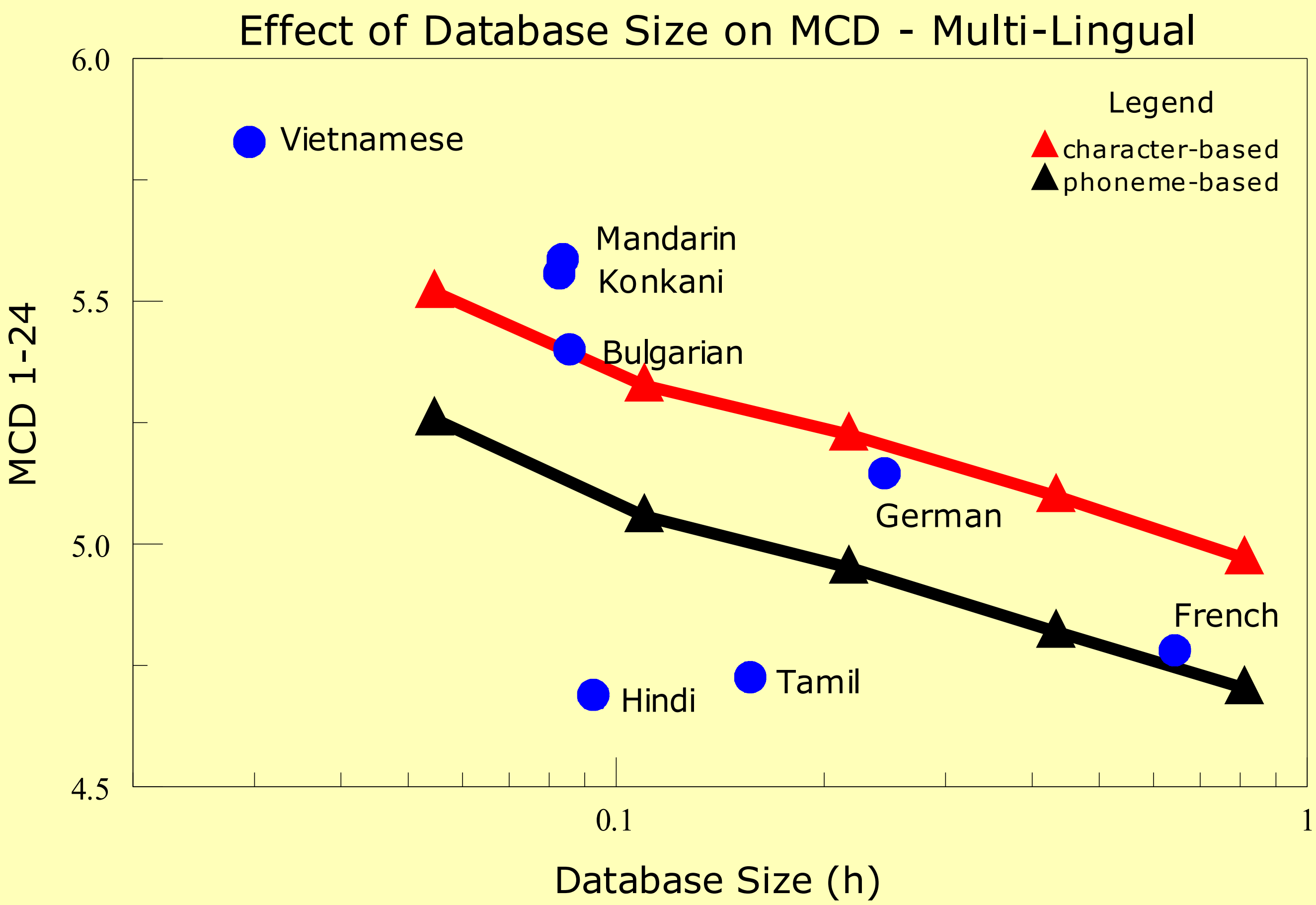
10 non-English test languages

- European
  - Bulgarian, French, German, Turkish
- Indian
  - Hindi, Konkani, Tamil, Telugu
- East Asian
  - Mandarin, Vietnamese

Evaluating non-English voices

- For a frame of reference, we need a *good* and a *bad* voice
  - Phoneme-based English is "*good*"
  - Grapheme-based English is "*bad*"
- Data covers 3m to 1h of speech
  - may be extrapolated to about 4h
- Non-English voices are from student lab projects

Non-English languages

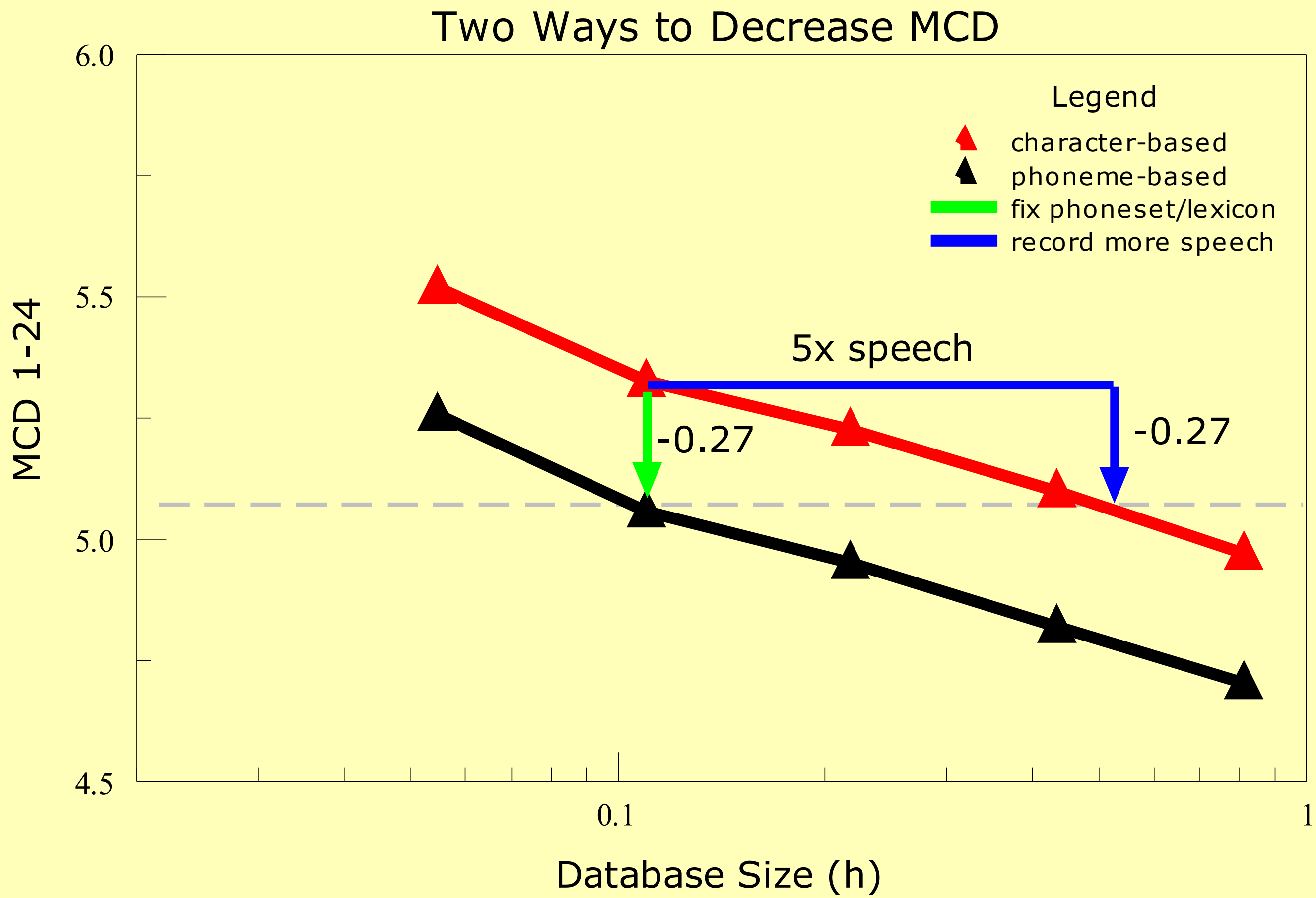


Characterizing voice quality

- Reference frame permits a quick assessment
  - French is in good shape
  - German could use lexicon improvements
  - Hindi and Tamil are good for their size
    - recommend: collect more speech
  - Bulgarian, Konkani and Mandarin need more speech and a better lexicon
  - Vietnamese voice had character set issues
    - resulted in only ¼ of the speech being used

More speech or a better lexicon?

- From the English MCD error curves
  - 5x the speech = fixing the phoneset + lexicon

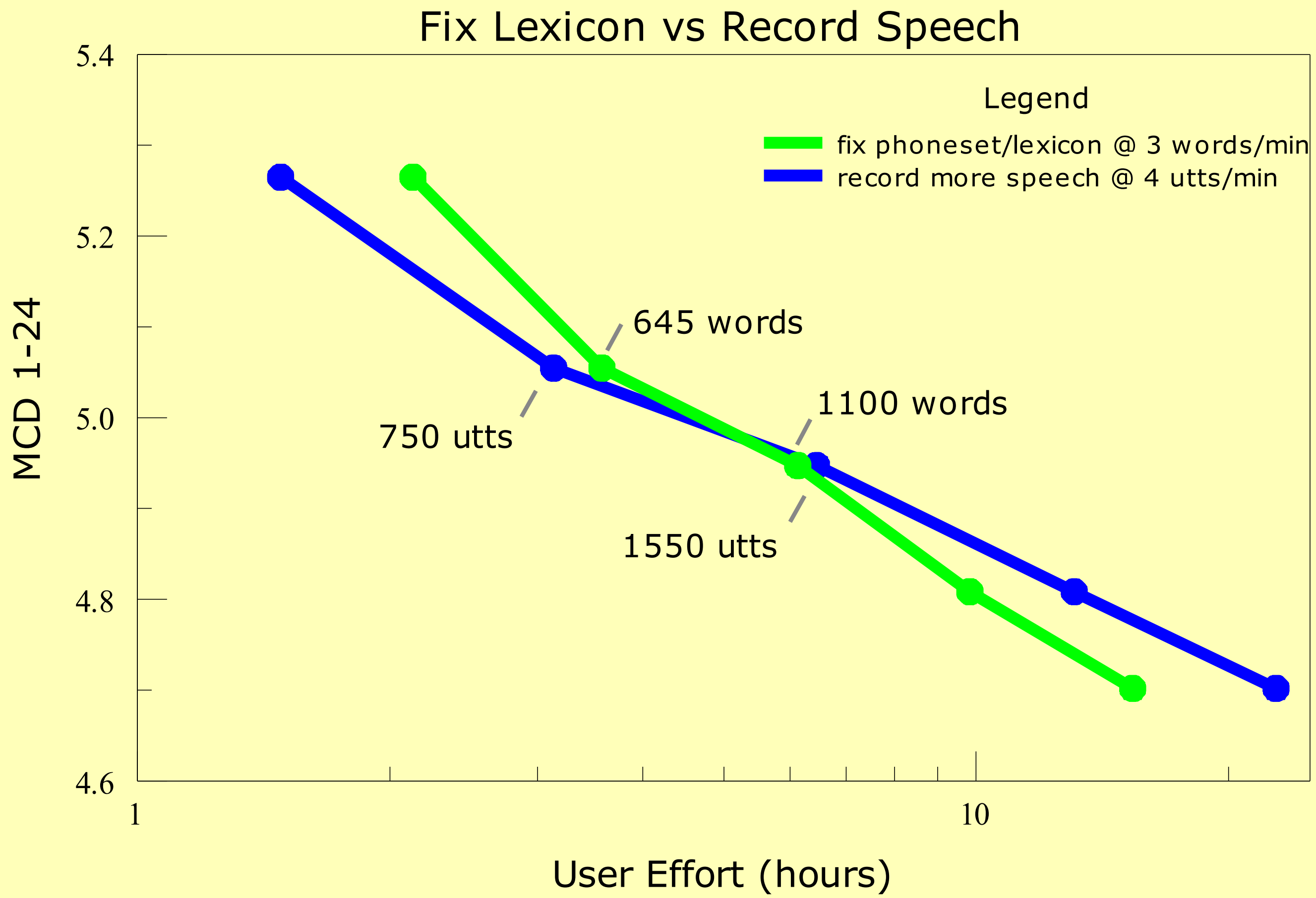


More speech or a better lexicon?

- Which is more time effective?
  - assume 3-4 sentence-length recordings per minute
  - assume 2-3 lexicon verifications per minute

- Answer
  - small database – record more speech
  - large database – work on the lexicon
  - the transition point is language-dependent
  - it also depends on the relative speed of recording and lexicon verification

More speech early, fix words later



Research Conclusions

- Language dependence
  1. Our language-dependent features are not critical
  2. Best stop value lies in 20-50 range, and is stable

- Measurement
  1. Cepstral distortion is useful quality measure
  2. Two "parallel lines" provide a frame of reference

- Efficiency
  1. Doubling speech reduces MCD by 0.12
  2. Adding lexicon to English reduces MCD by 0.27

Research Recommendations

- Human factors
  1. Interleave recording and lexicon work
     (too long on one task is mind-numbing)
  2. Emphasize recording early, lexical work later

- Future work
  1. Correlate MCD with listening tests
  2. Field testing with more users

- http://cmuspice.org