Text-to-Speech Synthesis

Prof. Dr. Tanja Schultz
(Folien und Synthesebeispiele von Alan W Black)

Universität Karlsruhe, Fakultät für Informatik
Speech Synthesis Process

Definition: **Speech synthesis** is the artificial production of human speech

Definition: A **Text-to-speech (TTS)** system converts written text (language) into speech

**Typically 3 Steps:**

- **Text Analysis**
  - From strings of characters to words

- **Linguistic Analysis**
  - From words to pronunciations and prosody

- **Waveform Synthesis**
  - From pronunciations to waveforms
Text Analysis

- Character Encodings
- Word Segmentation
- Numbers, Symbols, Non-Standard Words
  - Anything not directly in the lexicon
  - OOV or novel forms
  - Abbreviations, Letter sequences, Acronyms
- May require special rules
  - Train rules for homographs
  - Numbers, roman numerals

Text processing errors are the most common complaints in TTS
Text Analysis

- This is a pen.
- My cat who **lives** dangerously has nine **lives**.
- He stole $100 from the bank.
- He stole **1996** cattle on 25 Nov **1996**.
- He stole $100 million from the bank.
- It's 13 **St. Andrew St.** near the bank.
- Its a PIII 650Mhz, 128MB RAM, 13.6Gb SCSI, 24x cdrom and 19" monitor.
- My home page is http://www.geocities.com/awb/.
Text Analysis

“bunch of hacky rules”

But here, based on NSW Project from JHU Workshop '99
(NSW = Normalization of Non-Standard Words)

• **Splitter:**
  – domain independent tokenizer

• **token type classifier:**
  – number (ordinal/cardinal/year) – 1996 cattle vs year
  – homograph disambiguator – lives, St.
  – abbrev/letter sequence identifier – Mb, RAM, SCSI

• **token type expander:**
  – mostly deterministic
  – but abbreviation expander interesting

• **language model:**
  (optionally) choose best expansion
NSW models for specific domains

- Models for specific domains
- Standard text analyzers fail
- Can build models from labeled data

57 ST E/1st & 2nd Ave Huge

- Standard models
- Domain specific models
Synthesis Process

• Text Analysis:
  – from strings of characters to words

• Linguistic Analysis
  – from words to pronunciations and prosody

• Waveform synthesis
  – from pronunciations to waveforms
Lexicons and letter to sound rules

• Need big list of words:
  – often hardest requirement

• Sharing between dialects:
  – CSTR English dialect lexicon
    (Center for Speech Technology Research, Edinburgh)

• But the lexicon is never complete:
  – need out of vocabulary pronouncer
Bootstrapping Lexicons

- Find 250 most frequent words
  - Build lexical entries from them
  - Ensure letter coverage in base set
  - Build letter-to-sound (LTS) rules from this base set
- Select articles of text
- Synthesize each unknown word
  - Add correct words to base list
  - Correct incorrect words and add to base list
  - Rebuild LTS rules with larger list
  - Repeat process
Letter to sound rules

- Writing rules by hand is difficult
- We provide automatic process:
  - built from lexicon
  - provides phone string plus stress
  - tested on multiple languages

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Correct Letters</th>
<th>Correct Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>OALD</td>
<td>95.80%</td>
<td>74.56%</td>
</tr>
<tr>
<td>CMUDICT</td>
<td>91.99%</td>
<td>57.80%</td>
</tr>
<tr>
<td>BRULEX</td>
<td>99.00%</td>
<td>93.03%</td>
</tr>
<tr>
<td>DE-CELEX</td>
<td>98.79%</td>
<td>89.38%</td>
</tr>
<tr>
<td>Thai</td>
<td>95.60%</td>
<td>68.76%</td>
</tr>
</tbody>
</table>
Letter to sound rules: method

- Find alignments:

<table>
<thead>
<tr>
<th>c</th>
<th>h</th>
<th>e</th>
<th>c</th>
<th>k</th>
<th>e</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>ch</td>
<td>-</td>
<td>eh</td>
<td>-</td>
<td>k</td>
<td>-</td>
<td>t</td>
</tr>
</tbody>
</table>

- Using epsilon scattering and EM algorithm
- or hand specify possible letter-phone combinations

- Build CART models:
  - predict phone (or epsilon)
  - letter context (plus POS ...)

11
Classification Trees

Simple binary decision tree for height classification:

- T = tall,
- t = medium-tall,
- M = medium,
- m = medium-short,
- S = short

Goal: Predict the height of a new person

- Using decision tree to predict someone's height class by traversing the tree and answering the yes/no questions
- Choice and order of questions is designed knowledge-based
- Classification and Regression Trees (CART) provide an automatic, data-driven framework to construct the decision process
An example tree

For letter V: (Example: revving vs Romatov)

if (n.name is v) /* case vv */
   return _;

if (n.name is #) /* v at end of word */
   if (p.p.name is t) /* case t_v like in Romatov */
      return f;
   return v;

if (n.name is s)
   if (p.p.p.name is n)
      return f;
   return v;

return v;
Prosody

- Phrasing
  - chunking in speech
- Intonation
  - Accents and F0 generation
- Duration:
  - length of segments
- Power:
  - Loudness

Each has underlying “default”, and marked forms
Intonation marking

Text-to-Speech Synthesis
Intonation Marking

• Large pitch range (female)
• Authoritive since goes down at the end
  – News reader
• Emphasis for Finance H*
• Final has a raise – more information to come

• Female American newsreader from WBUR
• (Boston University Radio)
Intonation Examples

- Fixed durations, flat F0.
- Decline F0
- “hat” accents on stressed syllables
- Accents and end tones
- Statistically trained
Synthesis Processes

• Text Analysis:
  – from strings of characters to words
• Linguistic Analysis
  – from words to pronunciations and prosody
• **Waveform synthesis**
  – from pronunciations to waveforms
Waveform Synthesis

Concatenative synthesis:
- Random word/phrase concatenation
- Phone concatenation
- Diphone concatenation
- Sub-word unit selection
- Cluster based unit selection
Waveform Synthesis “diphone”

• All phone-phone transitions in language:
  – one occurrence of each
  – assume two-phone context is sufficient
  – may also include common consonant clusters
• Mid-phone is more stable so easier to join
• Inventory size: phones\(^2\)
• A “safe” way to build a synthesizer
Diphone example: (US English)

• Diphone schema 1348 words (1618 diphones):
  – includes consonant clusters and open vowels

• Prompts are:
  kal_0001 ("b–aa" "aa–b") (t aa b aa b aa)
  kal_0002 ("p–aa" "aa–p") (t aa p aa p aa)

• Automatic aligning alone:
  – around 15 words wrong
  – better than human labeler (first pass)
  – takes 30 minutes (vs 2 weeks)

• KAL voice created in 2 days
Unit selection synthesis

- Large representative database
- **Multiple** examples of each unit type
- Select “appropriate” units:
  - phonetically and prosodically
- Various unit selection algorithms:
  - target, concatenation cost
  - Cluster
  - phonological structure matching
- famed for being excellent
- famed for being incomprehensible
Cluster unit selection

- Cluster unit selection (Black&Taylor, Eurosp. 97)
- Decision tree partitioning on acoustic distance:
  - phone index (but diphone selection)
  - pitch synchronous melcep plus power and F0
  - indexed by phone and prosodic context
  - cf. Donovan 95
- Viterbi selection through candidates
  - cf. Hunt and Black 96, Iwahashi and Sagisaka 95
- Does it work?
  - good in parts
Making synthesis better

• Basic reading clean text in neutral form works
  – few applications require that though

• We need synthesis to be:
  • more flexible:
    – not just text to speech
  • more natural
    – so it's doesn't sound like a synthesizer
  • more efficient
    – easy to build new voices
    – synthesizes quickly on small machines
The boy saw the girl in the park with the telescope.
The boy saw the girl in the park with the telescope.
Some English first and then some Spanish.
Hola amigos.
Namaste.
Good morning My name is Stuart, which is spelled though some people pronounce it
stuart. My telephone number is 
2787.
I used to work in Buccleuch Place, but no one can pronounce that.
By the way, my telephone number is actually
http://att.com/sounds/touchtone.2.au
http://att.com/sounds/touchtone.7.au
http://att.com/sounds/touchtone.8.au
http://att.com/sounds/touchtone.7.au
What will the weather be like today in Boston?
It will be `<emph>rainy</emph>` today in Boston.

When will it rain in Boston?
It will be rainy `<emph>today</emph>` in Boston.

Where will it rain today?
It will be rainy today in `<emph>Boston</emph>`.
Making it efficient

- Adding new voices:
  - Need new voices and languages
  - tools to record label segment etc

- Synthesis must be fast
  - Dialog systems must respond in time
  - Recognition, dialog plus synthesis under 1 sec.
  - … and run on a handheld device or mobile phone
Summary – Essence of Synthesis

• Synthesis can be broken down into three parts:
  1. Text analysis:
     – expanding homographs, symbols, numbers etc
  2. Linguistic analysis:
     – pronunciation, letter to sound rules
     – prosody: phrasing, intonation and duration
  3. Waveform synthesis:
     – diphone voices:
     – unit selection voices
     – limited domain
• Not just text-to-speech:
  – concept to speech
  – markup
Build Your Own Synthetic Voices

The standard voice requires …

– A phone set
– Text analysis
– Pronunciations
– Prosodic modeling
– Waveform Generation

Plus something else that is difficult

– … and often language dependent
  • E.g. word segmentation in Chinese
  • Vowels in Arabic
  • Number declensions in Slavic languages
Multilingual Speech Synthesis

- Common technologies
  - (Diphone: too hard to record and label, coverage)
  - Unit selection is the standard method: *select appropriate sub-word units from large database of natural speech*
  - CONS: too much to record and label
  - Requires 200M per voice (single model across languages)

- New technology: Parametric Synthesis “clustergen (CG)”
  - HMM-generation based synthesis
  - Cluster units to form models, Generate from models
  - PRO: can work with little speech (10 minutes)
  - CONS: speech sounds buzzy, lacks natural prosody
  - Requires 2M per voice (single model across languages)
Unit Selection vs Parametric Synthesis

• Unit Selection: 🎤
  • large carefully labeled database (often 5000+ utts)
  • hard to speak 5000 good utterances > professionals
  • quality good when good examples available but rapid degradation when out-of-domain
  • little or no prosodic modifications
  • natural delivery; multiple speakers desired to model variability!!!

• Parametric Synthesis: 🎤
  • smaller less carefully labeled database
  • quality consistent
  • resynthesis requires vocoder, (buzzy)
  • can (must) control prosody
  • model size much smaller than Unit DB

Vocoder (voc (voice) and encoder) beruht auf der Zerlegung eines Eingangssignals in seine Frequenzbestandteile, der Übertragung dieser Bestandteile als Parametersatz, sowie der darauf folgenden Resynthese des Signals am Ziel auf der Basis der Parameter aus einem Rauschsignal.
HMM-Generation Synthesis

- NiTech’s HTS (Tokuda et al.)
  - Built as modification of HTK
  - FestVox build process exists
  - Hard coded feature values

- HMM-Generation Synthesis
  - Sub-phonetic features are modelled not as set of instances of units but parametric models
  - View these clusters as averages of instances

- High quality understandable speech (Bennett, 2005)
- Language Independence (Tokuda, 2002)
- Multilingual databases (GlobalPhone) within HMM-generation synthesis (Latorre & Furui, 2005)
ClusterGen

- New synthesis technique added to FestVox
  - Clustering technique for HMM-state sized segments
- Training data is HMM-based labeled speech
  - Labeling system included in FestVox
  - Janus RTk labels are used created by forced alignment
- CLUSTERGEN
  - Reversible (analysis/synthesis) parameterization of speech
  - MCEP analysis and MLSA filter for resynthesis (as in HTS)
  - 24-dim MCEP feature vectors
  - Clustered using wagon CART tree builder
  - Features for tree building are the articulatory features derived from GP IPA-based global phone inventory
  - Cluster optimization:
    - minimize sum of SD of each MCEP feature
    - weight by the number of samples in the cluster
Universal Sound Inventory & Data

Speech Production is independent from Language  ⇒  IPA

1) IPA-based Universal Sound Inventory

2) Each sound class is trained by data sharing

- Reduction from 485 to 162 sound classes
- *m,n,s,l* appear in all 12 languages
- *p,b,t,d,k,g,f* and *i,u,e,a,o* in almost all

GlobalPhone
- Use 10 languages
  Ch-Mandarin, Croatian, German, Japanese, Portuguese, Russian, Spanish, Swedish, Turkish + English (WSJ0)

- Use ASR global sound inventory
- Use IPA acoustic features
Clustering by CART

- Update to Wagon (Edinburgh Speech Tools)
- Clustering
  - F0 and MCEP, tested jointly and separately
  - Features for clustering (51): IPA articulatory features + other phonetic, syllable, phrasal context
- Training Output - Three models:
  - Spectral (MCEP) CART tree
  - F0 CART tree
  - Duration CART tree
- F0 model:
  - Smooth extracted F0 through all speech (i.e. unvoiced regions get F0 values)
  - Chose voicing at runtime phonetically
FestVox CLUSTERGEN Synthesizer

- Prompt transcriptions, Waveform files (well recorded)
- Labeling
  - Used CI models and forced alignment (JRTk – monolingual ASR)
- Parameter extraction:
  - (HTS’s) MCEP/MLSA filter for resynthesis
  - F0 extraction
- Clustering
  - Wagon vector clustering for each HMM-state name
- ClusterGen Synthesis:
  - Generate phoneme strings (as before)
  - For each phone:
    - Find HMM-state names: ah_1, ah_2, ah_3
  - Predict duration of each
    - Create empty MCEP vector to fill duration
    - Predict MCEP and F0 values from corresponding cluster trees
    - Use MLSA filter to regenerate speech
Measuring Quality

Mean Mel Cepstral Distortion (MCD) over test set (smaller is better)

\[
10 / \ln 10 \sqrt{2 \sum_{d=1}^{24} \left( mc_d^{(t)} - mc_d^{(e)} \right)^2}
\]

Measured on a Cross evaluation set

MCD: Voice Conversion ranges 4.5-6.0
MCD: ClusterGen scores 5.0-8.0
Manual speaker selection

⇒ For all languages monolingual TTS performs best
⇒ Multilingual Models perform well …
    … only if knowledge about language is preserved (Multi+)
    (only small amount of sharing actually happens)
Speaker Clustering

Hierarchical Bottom-up Clustering

BIC stopping criterion

$$\Delta BIC = BIC(M_C) - BIC(M_{A,B})$$

$$BIC(M) = \log L(X | \mu, \Sigma) - \frac{\lambda}{2} V(M) \log N$$

$$M_{A,B}: X_A \sim N(\mu_A, \Sigma_A) \quad X_B \sim N(\mu_B, \Sigma_B)$$

$$M_C: X_C \sim N(\mu, \Sigma)$$

TGMM-GLR distance

$$D(S_A, S_B) = -\log \frac{P(X_{A \cup B} | \theta_{A \cup B})}{P(X_A | \theta_A)P(X_B | \theta_B)}$$

$$= \log \frac{P(X_A | \theta_A)P(X_B | \theta_B)}{P(X_C | \theta_C)}$$
Manual vs Clustered Speaker Selection

- Selecting similar speakers helps for both Mono and Multi
- Multi benefits more as expected: similarity more important
- Large variation across languages (label quality? Data size?)
Conclusion & Future Work on ML-TTS

- ClusterGen allows for much smaller voices
  - Multilingual voice for unit selection: 200Mb
  - Multilingual voice for ClusterGen: 2Mb
- Preserving language information (Multi+) helps
- Selecting similar speakers helps
- All voices are understandable (no formal tests!)

Future Work:
- ClusterGen is *very* young
  - No treatment of dynamics yet
  - Multilingual, Multispeaker DBS
  - Grapheme Based models
  - Voice/Style conversion: model interpolation
  - Signal reconstruction: proper residual modeling
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)
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

Grapheme versus Phoneme-based Voices

Legend
- grapheme based
- phoneme based

improvement from good phone set and lexicon

MCD 1-24

0.27

Wagon Stop Value
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
Effect of Database Size on MCD - Multi-Lingual

- **Legend**
  - ▲: character-based
  - ▼: phoneme-based

- **Languages**
  - Vietnamese
  - Mandarin
  - Konkani
  - Bulgarian
  - German
  - French
  - Vietnamese
  - Hindi
  - Tamil

- **Database Sizes**
  - 0.1
  - 1
  - 3.5
  - 4.0
  - 4.5
  - 5.0
  - 5.5
  - 6.0

- **MCD 1-24**

- **Database Size (h)**

- **Non-English languages**

Text-to-Speech Synthesis
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 $\frac{1}{4}$ of the speech being used
More speech or a better lexicon?

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

Two Ways to Decrease MCD

Legend
- character-based
- phoneme-based
- fix phoneset/lexicon
- record more speech

5x speech

Database Size (h)

MCD 1-24
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

Fix Lexicon vs Record Speech

Legend
- fix phoneset/lexicon @ 3 words/min
- record more speech @ 4 utts/min

User Effort (hours)

MCD 1-24

- 645 words
- 750 utts
- 1100 words
- 1550 utts

Text-to-Speech Synthesis

Cognitive Systems Lab

Karlsruhe Institute of Technology
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
Focus on TTS

• Main research questions
  1. To what extent is language-dependent expertise required of the user?
  2. To improve the synthesizer, what is the most efficient use of the user's time?
  3. How can we measure the user's progress?
Research question in detail

• Language dependence
  1. Which features matter the most in CART tree training? Are language-dependent features critical?
  2. What is the best 'stop value' for training?

• Measurement
  1. Can an objective measure be used to estimate the quality of a voice, in any language?
  2. Can this information motivate and inform the user?

• Efficiency
  1. Rate of improvement as more speech is recorded?
  2. Rate of improvement as the lexicon is expanded and corrected?
TTS overview

- "welcome"
  - lexicon or LTS

- W EH L K AH M

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 EH₁ state we use name feats
  - prev states: ... AY₃ W₁ W₂ W₃
  - next states: EH₂ EH₃ L₁ L₂ ...
  - prev phones: # HH AY W
  - next phones: L K AH M
More CART tree features

• Four categories of training features
  1. names: phoneme and HMM state context
  2. position: e.g., number of frames from beginning of state, percentage in from beginning
  3. IPA: International Phonetic Association features, based on phoneme set
  4. 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 $\nu$ = sequence of 25-D cepstral frames, 5 ms step
Effect of isolated feature classes

Feature Classes in Isolation

Legend

- ling
- IPA
- names
- posn

Wagon Stop Value

MCD 1-24

better

worse
Effect of combined feature classes

Cummulative Feature Classes

Legend
- names
- names + posn
- names + posn + IPA
- + linguistic (hidden)

MCD 1-24

Wagon Stop Value

baseline
over-fit
under-fit

+names
+posn

optimal range
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...

<table>
<thead>
<tr>
<th>Feature class</th>
<th>Features</th>
<th>Lang dep.</th>
<th>Δ MCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>no CART trees</td>
<td>0</td>
<td>no</td>
<td>baseline</td>
</tr>
<tr>
<td>name symbolics</td>
<td>16</td>
<td>no</td>
<td>-0.452</td>
</tr>
<tr>
<td>position values</td>
<td>7</td>
<td>no</td>
<td>-0.402</td>
</tr>
<tr>
<td>IPA symbolics</td>
<td>72</td>
<td>yes</td>
<td>-0.001</td>
</tr>
<tr>
<td>linguistic sym.</td>
<td>14</td>
<td>yes</td>
<td>+0.004</td>
</tr>
</tbody>
</table>
Effect of database size

• Doubling speech reduces MCD by $0.12 \pm 0.02$
  – 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

The diagram shows the effect of database size on MCD curves. The optimal stop value is fairly stable across different database sizes tested on 10% heldout.

Legend:
- 1/16 hour
- 1/8 hour
- 1/4 hour
- 1/2 hour
- 1 hour

Tested on 10% heldout.
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
Trial experiences

• Jan-May 2007 – 11733 lab course
  – hands on laboratory in building speech-to-speech translation system using SPICE tools
  – 11 students covering Bulgarian, Chinese, German, Hindi, Kankani, Thai, Turkish

• Lessons learned
  – phoneme selection too hard
  – making corrections was awkward
    “don't bother me when the prediction is correct”
  – want synthesized pronunciation in their voice
  – need guidance on when to attempt voice build
  – need integrated evaluation and testing
1-D Illustration of MCD

MCD distortion is average distance between curves

-1.0
0.0
1.0
2.0
3.0

Legend
original
synthesized

Frame Number
0 100 200 300 400 500
first cepstral dimension
0.0
1.0
2.0
3.0
4.0
5.0

original
synthesized
More speech early, fix words later

- e.g. third line
  311 utts + fixed lexicon = 1546 utts, grapheme-based
  387 min to record 1546 utts only (fast case)
  368 min to record ¼ hour and fix lexicon (fast case)
Phonemes are hard to identify

- linguists spend years in careful analysis
- non-linguists are baffled
Phonemes are hard to use

- Spelling usually different from pronunciation
  - getting a word right is hard
  - keeping it consistent is even harder

Contributions

LexLearner

- system selects new word
- or correct word by hand
- verify and submit word

CMU SPICE

Text and prompt selection (help)
Audio collection (help)
Phoneme selection
Grapheme-to-phoneme rules (help)
Lexicon pronunciation creation (help)
Build acoustic model (help)
Build language model (help)

Phoneme labels for your language:
P B T D K G M N R F W S Z SH ZH H L I Y U E O A YA YU YZ DZ TCH
Good G2P rules ease lexicon task

Pronunciation Coverage (Emille, WSJ)

- 60% coverage in 400 words
- 50% coverage in 200 words
- 40% coverage in 100 words

Tokens covered by Words (%) vs Word Rank

English

Hindi