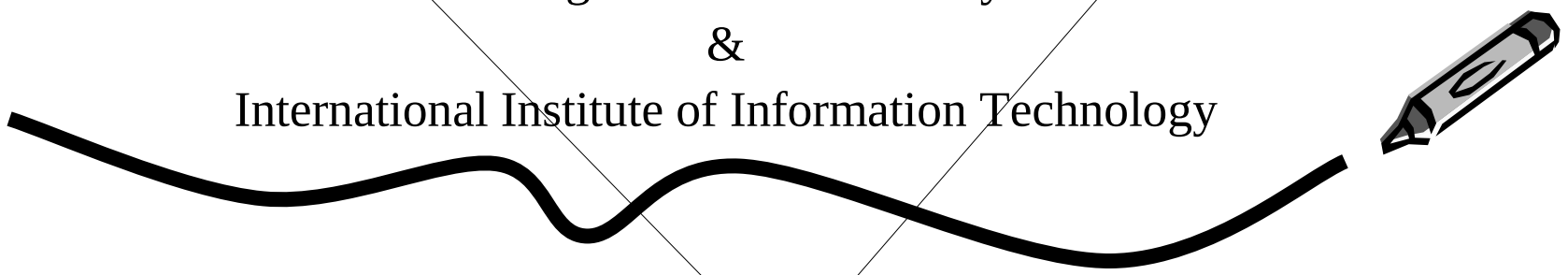
# Grapheme Based Speech Synthesis and Speech Recognition

S.P. Kishore
skishore@cs.cmu.edu

Carnegie Mellon University
&
International Institute of Information Technology



# Focus Papers

- Alan W Black and Ariadana Font Llitjos, “Unit Selection With out Phoneme Set”, IEEE TTS Workshop 2002, Santa Monica, CA
- Mirjam Killer, Sebastian Stuker, Tanja Schultz, “Grapheme based Speech Recognition”, Eurospeech, Geneva, 2003.

# Speech Synthesis without Phoneme Set

- Why without phoneme set?
- Increasing need for speech synthesis and speech recognition in new unsupported languages.
  - Availability of less phonetic knowledge in the new languages
  - Researchers/developers may not have (or access to) language expertise
  - Native speakers may not be consciously aware of the phonetic knowledge
- How hard/easy for well studied languages? –
  - An appropriate phoneme set for a well studied language may not be easy.
  - Orthography and phonetics may not have one to one relationship

# Experiment in Spanish language

- Nature of the language: Written system and phonology is relatively close, but not one-to-one
- Letter set as phoneme set
  - 26 standard English letters
  - Accented characters a', e', i', o', u' and n'
- Pronunciation:
  - Word into characters (list of phones)
  - No vowel/consonant information available!
  - Each word is coded as a single syllable
  - Numbers?: Expanded into complete words using knowledge base

# Labeling

- Typical Process using DTW (in Festvox):
  - Phone set and duration information is available
  - Prompts are generated
  - Use this acoustic and duration information to do a DTW on the uttered sentences
- If no Phone Set?
  - Labeling using acoustic models of speech recognition systems (Sphinx Tools)
  - Acoustic models built using letter as phone names
  - Once the models are trained, segmental information could be obtained.

# How Good These Systems are?

- Confirm with different pronunciation of letters in different context
  - Letter context and position information in the word is useful
  - Ex: **c**asa → /**k** a s a/ (house)
  - **c**esa → /**th** e s a/ (stop)
  - **c**ine → /**th** i n e/ (cinema)
  - **c**osa → /**k** o s a/ (thing)
- Could capture dialect differences

| | Castillian | Colombian |
|---|---|---|
| – cesa | → /th e s a/ | /s e s a/ |

- Synthesis Quality: Results show good rating for 90% of words.

# Pros and Cons..

- Overcome the effects of using one language/dialect phone set onto another (ex: Pronunciation of Scottish English speaker does not match with US English lexicon!!)
- Does not require linguistically knowledgeable speakers of the language
- May not be easy to specify/formulate fine distinctions
- Letter to sound rules may not be easy

- Requires sufficient data for the model to get trained

# Speech Recognition

- Pronunciation Dictionary: Core component
- Each lexicon entry is mapped to sequence of sub word units (phonemes)
- Accuracy of ASR systems heavily depend on consistency and accuracy of pronunciation dictionary

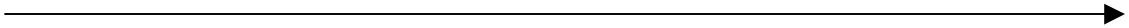
- For new languages, automated generation of pronunciation dictionaries rule-based or statistical based approaches.
  - Dictionaries – hand crafting is time consuming
  - Non-accurate dictionaries degrades the ASR performance

# Grapheme Vs Phoneme ASR

- Phoneme based systems
  - 3 – state HMM with 3000 triphone models
  - 32 Gaussians for each HMM state
  - Linguistically motivated questions to cluster the polyphonic decision tree
- Grapheme based systems
  - As in the case of phoneme, modeled by 3-state HMM
  - Pronunciation dictionaries: - split the word into characters
- Decision trees for context dependent modeling

# Performance

| | WER | | |
|---|---|---|---|
| Language | English | German | Spanish |
| Phoneme | 12.7% | 17.7% | 24.5% |
| Grapheme | 19.1% | 17.0% | 26.8% |



Grapheme-phoneme Correspondence

# Context Width of the Models

- A context width of one (C-1) leads to tri grapheme system
- A hybrid tri grapheme system in which question and model context windows are different.

| Language | C-1 | C-1 Q-2 | C-2 | C-2 Q-3 | C-3 |
|---|---|---|---|---|---|
| English | 19.1% | 19.8% | 21.7% | 22.4% | 23.6% |
| German | 18.1% | 17.0% | 18.4% | 18.7% | 18.7% |
| Spanish | 27.0% | 26.8% | 28.8% | 28.2% | 31.4% |

# Other Ideas

- Question generation to group poly-grapheme into a limited number of clusters.
- Multilingual Grapheme based Recognition
  - Rapid adaptation to new languages
  - Similar to multilingual phoneme based speech recognition

• • • •

- Questions and Discussion
- Future directions on the topic.