ADVANCES IN ALPHADIGIT RECOGNITION USING SYLLABLES

Jonathan Hamaker, Aravind Ganapathiraju, Joseph Picone

Institute for Signal and Information Processing
Mississippi State University
{hamaker, ganapath, picone}@isip.msstate.edu
http://www.isip.msstate.edu

John J. Godfrey

Personal Systems Laboratory
Texas Instruments Inc.
godfrey@csc.ti.com
http://www.ti.com

Motivation

- Applications in automated telephony, information retrieval and security
- Alphadigit performance saturated at 10% WER
- Preliminary results on Switchboard (WS'97)
- Lower complexity than traditional triphone-based systems

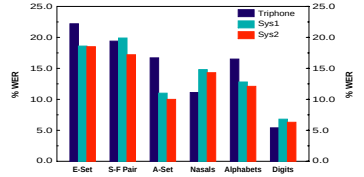
OGI Alphadigits

- Telephone database collected digitally using a T1 interface to the telephone network
- 3000 subjects in the corpus
- 19 or 29 alphanumeric strings per speaker
- All strings were exactly six words in length ("8 H A 8 B H", "8 W R W 8 E")
- 1102 unique prompting strings
- Balanced phonetic context of bigrams

Triphone System

- A word-internal triphone system and a cross-word triphone system
- 3-state left-to-right models without skips
- 12 Gaussian mixture components per state
- Cross-word system had 25202 virtual triphones, 3225 real triphones, 2045 states.
- Performance: Cross-word — 12.2%
  Word-internal — 13.4%

Error Modalities

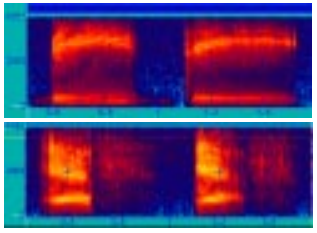


- The syllable system does better than triphones on A-set, E-set and S-F pair
- Syllables achieve greatest gains in alphabets: 12.1% WER compared to 16.5% for triphones

Common Problems

E-Set "b d"

S-F Pair "f s"



- Common error modalities are: S-F Pair, E-Set, Nasals and A-Set

Database Partitioning

	Number of Speakers / Utterances		
	Male	Female	Children
Training	1064 / 24611	1150 / 26405	22 / 500
Dev Test	355 / 8200	384 / 8867	8 / 188
Eval	71 / 1582	77 / 1710	2 / 37

- Standard partitioning for an SI evaluation
- Balances percentage of males, females and children across all sets

Syllable Systems

- Unique number of states per model
- Sys1:
  - No. States: 2 x median duration
  - Complexity: 42 syllables / 900 states
  - WER: 11.1%
- Sys2:
  - Max. states/syllable: 20 states
  - Complexity: 42 syllables / 702 states
  - WER: 10.4%

Conclusions

- Syllables reduced WER by 14.8% and yielded a lower complexity system
- No explicit pronunciation modeling required
- Model topology needs to be data-driven
- Alphadigits are an adequate experimental framework for investigating syllable models
- Greater gains expected for conversational speech applications

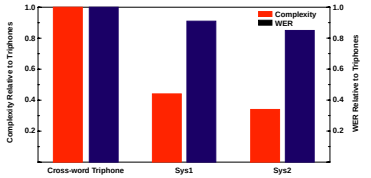
Previous Approaches

- Detailed phonemic modeling (Spanias and Loizou, 1996)
- Modeling of onsets, spectral transitions and glottal stops
- WER of 15% on speaker-independent (SI) alphabet task
- Improved feature representation using spectral warping (Mashao, 1996)
- WER of 8.2% on connected alphadigits (SI)

Why Syllables?

- Triphone durations are too small
- Unsuitable for integration of spectral and temporal dependencies
- Syllables provide larger acoustic context useful for modeling coarticulation
- Syllables yield dramatic reduction in system complexity
- Syllables better represent human perception

System Comparison



- Best syllable system decreases WER by 14.8%
- Complexity of the best syllable system is 66% less than the triphone system

Future Work

- Use information theoretic measures for optimal topology — Bayesian Information Criterion or Minimum Description Length
- Limited use of context-dependent syllable modeling and multi-syllable phrases
- State-tying with syllables
- Explore discriminant classification approaches such as Support Vector Machines