RESEGMENTATION OF SWITCHBOARD

Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, Joseph Picone

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

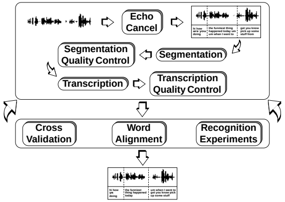
Introduction to SWITCHBOARD

- Challenging and popular LVCSR benchmark
- Spontaneous telephone conversations
- 240 hours, 2430 conversations, 3+ million words, 500+ speakers (male and female)
- Low bandwidth, channel noise, echo
- Speaking rates, dialects, coarticulation, speaking styles, accents, dysfluencies
- Poor quality acoustic models, large mismatch

Motivation

- Reduce acoustic model mismatch
- Segmentation and transcription must capture both acoustic and linguistic properties
- Automatic (energy-based) segmentation — unnatural breakpoints
- Linguistic structure-based segmentation — corrupted acoustic context
- Dysfluencies make transcription difficult (Current LDC transcription WER ~ 8%)

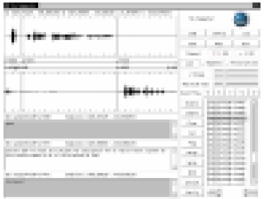
Approach


Echo Cancel
hi how are you doing
the funniest thing happened today um um when I went to
get you know pick up some stuff from
Segmentation
Segmentation Quality Control
Transcription
Transcription Quality Control
Cross Validation
Word Alignment
Recognition Experiments
hi how ya doing
the funniest thing happened today
um when I went to get you know pick up some stuff

Guidelines

- Segment boundaries with at least 1 sec of silence between speech
- Segment along phrase / sentence / train-of-thought boundaries
- Merge utterances split at counterintuitive points (e.g. middle of sentence)
- Limit maximum utterance duration to 15 sec
- Fix transcriptions taking into account dysfluencies and capitalization issues

Segmentation Tool

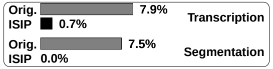


Issues and Concerns

- Large number of dysfluencies (pauses, laughter, partially pronounced words etc.)
- Affirmative statements (yes/no) and pause fillers (um/hmm) cover ~ 30% of utterances
- Marking boundaries near noise or echo
- Consistency in capitalization (“I” vs “i”) and handling proper nouns
- Marking asides, background noise / music and background speech

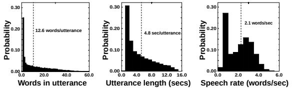
Cross-Validation

- All validators segment / transcribe the same conversation
- Adjudicated reference transcription
- Word alignment review will further reduce error rate



Cross-Validation Word Error Rates

The New SWITCHBOARD



- Segmentation and transcription rate 20xRT
- Monosyllabic words constitute 53% of data on WS’97 subset (down from 67%)
- Lexicon updates — partial words, laughter words, alternate pronunciations

Effect on Recognition

- Adapt existing acoustic models to resegmented speech data
- 20 hours training data (27500 utterances) including silence
- Word-internal triphone system to bootstrap seed models (HTK)
- 4 passes of re-estimation
- Lattice rescoring on WS’97 dev test set

Results

Error Rate	ISIP	WS’97
Total WER	47.9%	49.8%
Correct words	55.8%	53.1%
Substitutions	31.6%	32.2%
Deletions	12.6%	14.8%
Insertions	3.7%	2.9%

- 63% of total errors on monosyllabic words (down from 71%)
- Reduction in substitution and deletion errors

Analysis

- 1.9% absolute improvement in WER
- Monosyllabic words are the principal factor in error analysis
- Performance improvement proportional to better modeling of monosyllabic words
- Acoustically “complete” transcriptions help in improved acoustic modeling
- Longer utterance transcriptions facilitate LM application

Conclusions

- Uniformity and accuracy are critical for the quality of training segmentation and transcriptions
- Segmentation at natural boundaries allows better acoustic models
- Dysfluencies pose significant challenges to accurate transcription
- Acoustic models trained on corrected SWB data will result in major improvements in WER (e.g. 2% absolute improvement on adaptation)