Contact: matthias.sperber@kit.edu

Efficient Speech Transcription Through Respeaking

Matthias Sperber[1,3], Graham Neubig[2], Christian Fügen[3], Satoshi Nakamura[2], Alex Waibel[1]

[1] Karlsruhe Institute of Technology, Germany; [2] Nara Institute of Science and Technology, Japan; [3] Mobile Technologies GmbH, Germany

Introduction

- A respeaking method that can make post-correction…
  - **more efficient** than with typing, and…
  - **less demanding** than with traditional respeaking methods.

Approach

Initial step: speaker enrollment

→ Acoustic model adaptation using maximum likelihood linear regression

Step 1: Create automatic transcript

→ Should be as accurate as possible to reduce supervision effort, and make confidence filtering effective

Step 2: Segment into sentence-like units

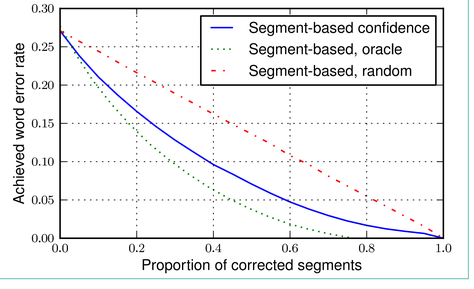
Desirable properties:

- Sentence-like units (natural to respeak)
- Not too short (→ language model context needed for respeaking recognition)
- Not too long (→ allow fine-grained confidence filtering; obtain more perfect segments so we can skip respeaking)

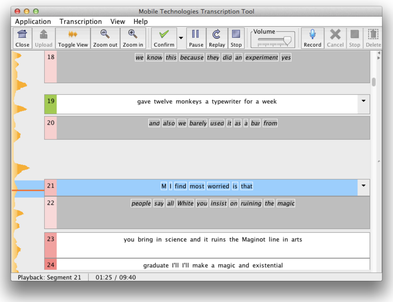
Approach: use [Matusov+06]'s segmentation method (log-linear feature combination including prosodic-, language-model-, and other features)

Step 3: Confidence annotation & filtering

- Segment confidence = mean word-posterior
- Only supervise segments with confidence below threshold



Step 4: Respeaking of selected segments



Important: skip respeaking when segment already correct initially (respeaking can only make it worse!)

Step 5: Combine respoken and initial transcripts

- Different recognition errors for original speech and respeaking
- Align both using phone similarity, choose most confident word at each position

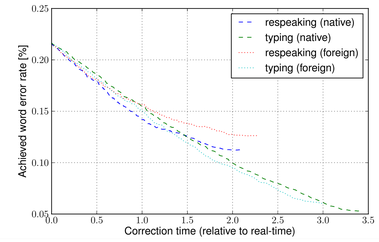
Experiments & Results

- Respeaking task on TED talks
- 1 native speaker, 1 foreign speaker
- Both: inexperienced respeakers, fast typists
- Typed and respoke confidence-ordered segments

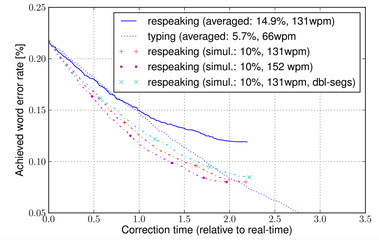
Input speed (wpm)	
Typing	58~61
Respeaking	83~97

Recognition Accuracy (WER)				
	Original	Respoken	Combined	Typed
Correct all	21.7	14.9	13.1	5.7
Skip correct		12.3	**11.9**	**5.7**

Respeaking vs. typing



Simulations



Respeaking

- Faster (WER 21.7 → 11.9, ~2x real-time)

Typing

- More accurate (WER 21.7 → 5.7, ~3.5x real-time

Typing faster if segment WER < 5% (respeaking is for whole segment, whereas only parts that contain errors need to be re-typed)

Use of segmentation

+ Makes respeaking much easier (cheaper!) than respeaking everything without break
− Causes 1.8% additional errors (due to inaccurately aligned segment breaks)

- Avg. segment length: 8.6 words (which is a reasonable trade-off: simulating double segment length decreased overall efficiency)

Confidence filtering

+ Greatly helps efficiency
− Can be confusing to annotator ("jump" through the speech) → unsupervised part of transcript should be displayed as text to convey context

- Simulating better (but still realistic) WER/speed greatly boosts efficiency
- Hypothesis combination effective
- **Bottom line: provided reasonable respeaking WER, respeaking is more efficient (except near-perfect segments that should be typed)**