Role Annotated Speech Recognition for Conversational Interactions

Nikolaos Flemotomos[1], Zhuohao Chen[1], David C. Atkins[2], Shrikanth Narayanan[1]

[1]Signal Analysis and Interpretation Lab (SAIL), University of Southern California
[2]Department of Psychiatry and Behavioral Sciences, University of Washington

USC University of Southern California

SAIL

Motivation & Idea

Automatic rich transcription when speakers have roles:

- Automatic Speech Recognition (ASR)
- Speaker Diarization
- Speaker Role Recognition (SRR)

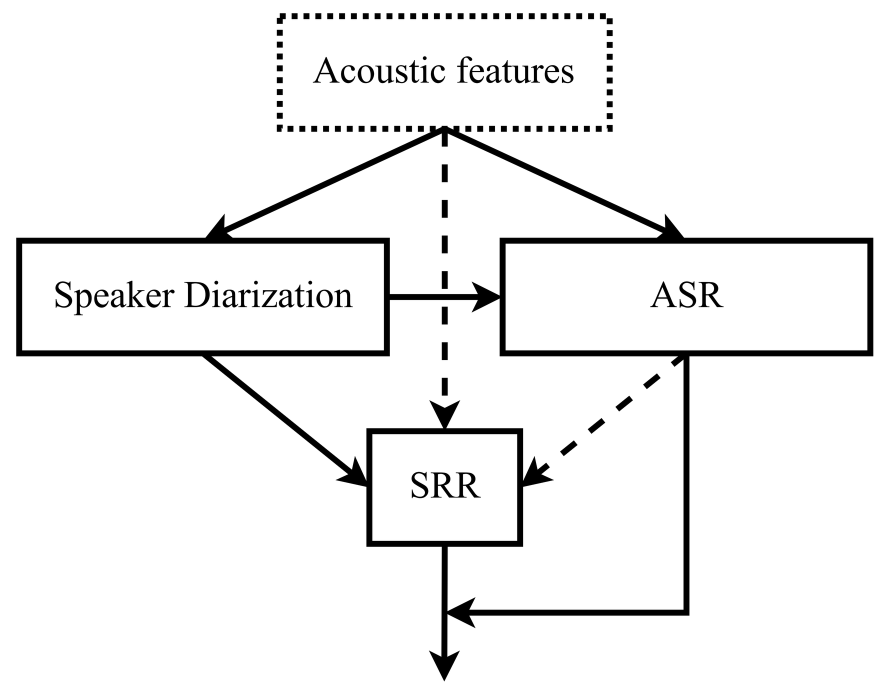


Figure 1: Traditional approach for automatic rich transcription.

Problem: error propagation
Solution: end-to-end system
⇒ *Role-Annotated Speech Recognition (RASR)*

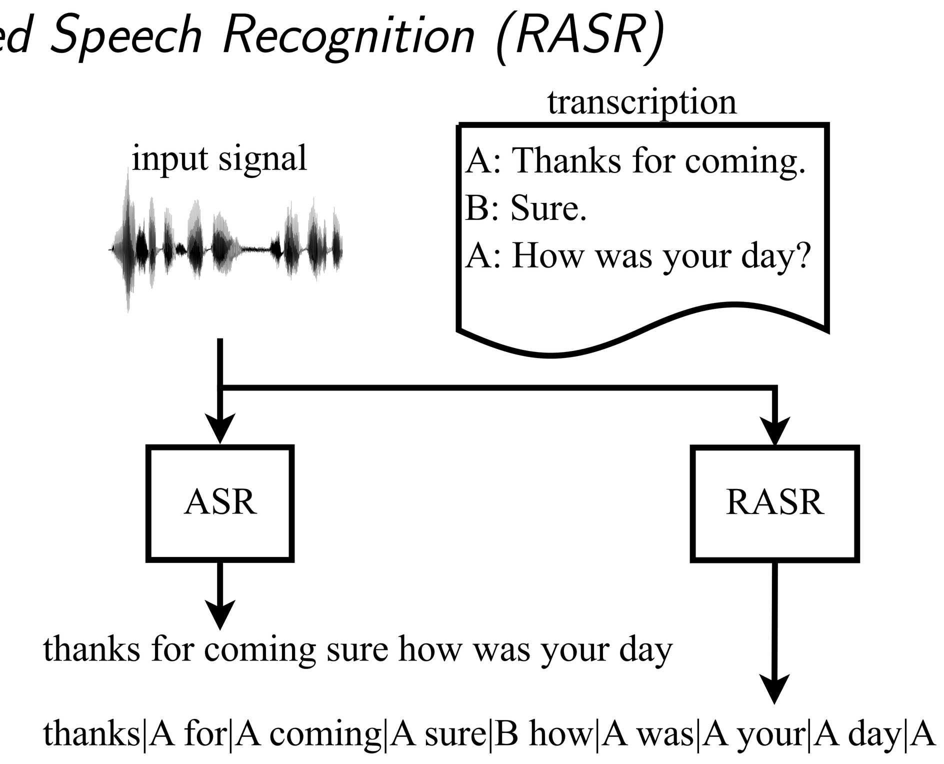
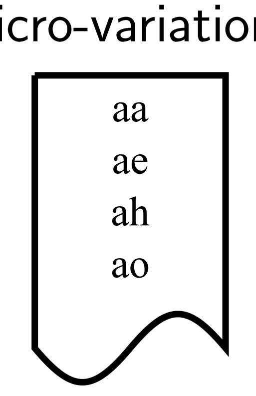


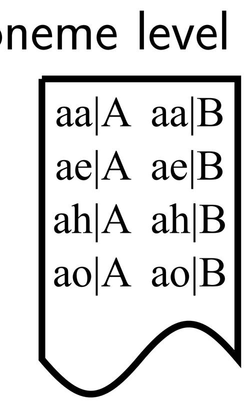
Figure 2: ASR vs. RASR for input segment with 2 roles, A and B.

Method

- Extend the phoneme set to include role annotations.
  ⇒ capture micro-variations between roles at the phoneme level

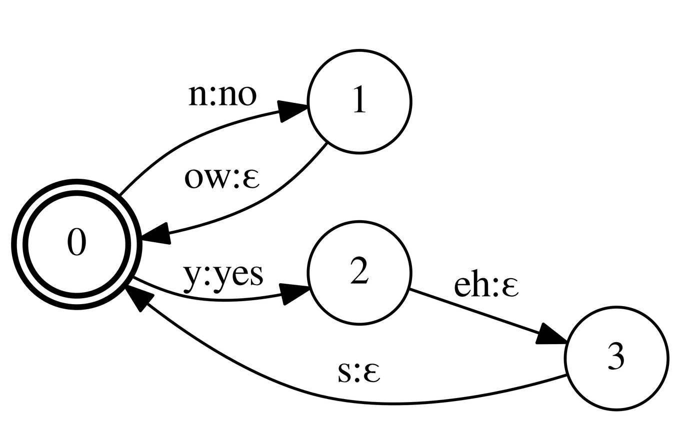


(a) Original phoneme set.

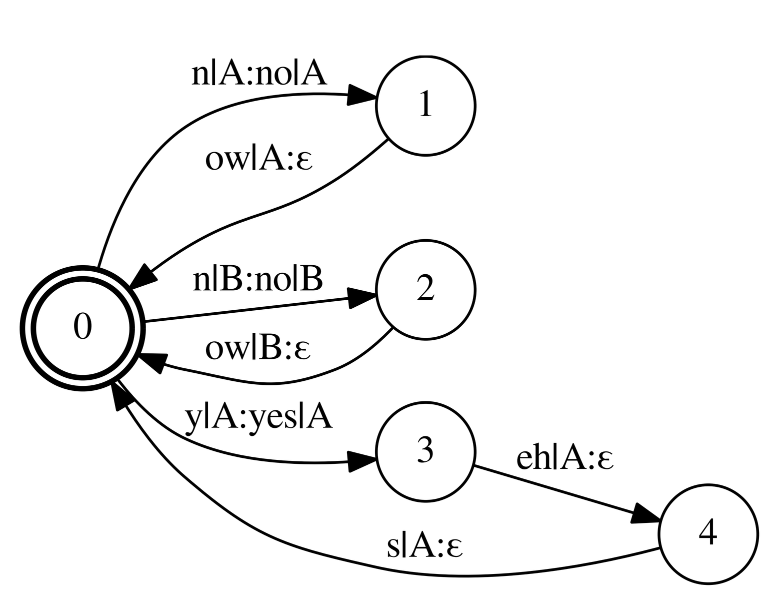


(b) Role-annotated phoneme set.

- Build the role-annotated lexicon.

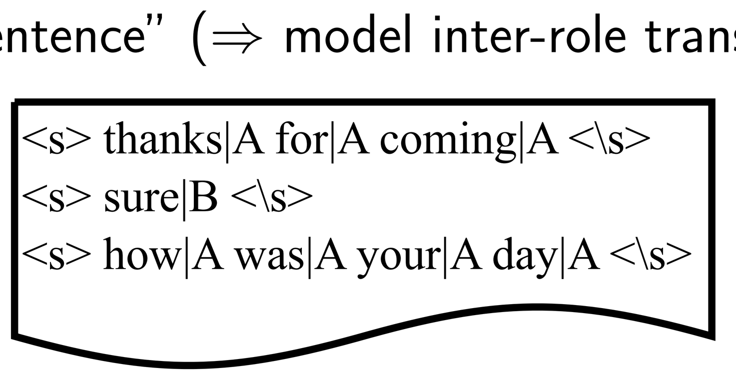


(a) Non-annotated lexicon.

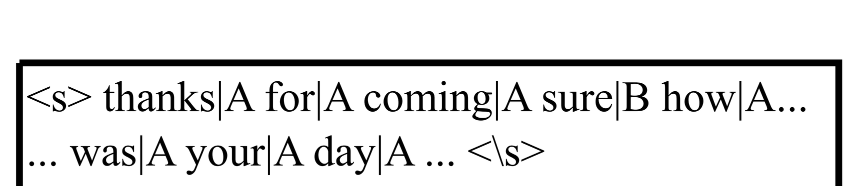


(b) Annotated lexicon.

- Train the acoustic model. Role-annotated versions of the same phoneme share the same root in the phonetic decision tree?
- Train the language model on role-annotated corpus. Treat each session as a "sentence" (⇒ model inter-role transitions)?



(a) One sentence per speaker turn.



(b) One sentence per session.

Speaker Normalization for RASR

- Speaker adaptation is an essential element of ASR. (CMN, SAT, i-vectors, etc)
- But in RASR
  a) the assumption of one speaker per segment does not hold,
  b) speaker variability is helpful to deduce roles.
  - do not use SAT (alignments for DNN based on LDA-MLLT)
  - online CMN (speakers do not change fast)
  - provide i-vectors for DNN training (let the network decide)

Experiments

Dataset

dyadic interactions (therapist-T vs. client-C) in psychotherapy

	#sessions	dur-T	dur-C
train	74	22.40 h	18.96 h
test	69	17.76 h	14.70 h

Procedure

- Force-align both training and test sessions.
- Segment training sessions according to manually derived speaker turns & test sessions according to whether the pause between 2 words is longer than 1 sec.
- Train RASR with the two role annotations (T, C). Online CMN and i-vector extraction using a 2-sec history window.

Evaluation metrics

- ASR performance: Word Error Rate (WER)
  - discard role annotations
- Diarization & SRR performance: Role Error Rate (RER)
  - use the alignments of the output to extract turn boundaries
  - computation similar with Diarizaiton Error Rate (DER)

Results

- Using RASR to perform jointly ASR and Diarization (& SRR)

	conc share	conc no-share	no-conc share	no-conc no-share
CMN	**39.74**	41.32	39.86	41.27
no-CMN	41.84	42.63	41.47	43.82

(a) RER (%)

	conc share	conc no-share	no-conc share	no-conc no-share
CMN	58.82	61.47	**58.78**	61.37
no-CMN	63.64	65.07	63.45	65.13

(b) WER (%)

The annotated versions of the same phoneme may or may not share the same root of the phonetic decision trees (*share* vs. *no-share*) and the LM may be trained on a corpus which contains all the speaker turns independently (*no-conc*) or concatenated per session (*conc*).

- Pre-trained tools for Diarization and ASR
  DER = 39.61% (LIUM SpkDiarization)
  WER = 41.27% (Kaldi ASpIRE model)
- In-domain training and RASR-based normalization for ASR
  WER = 54.21%

Challenges - Future Work

- Conflicting goals of ASR and Diarization/SRR
  ⇒ suitable feature engineering for the hybrid task of RASR
- adaptation of pre-trained ASR models to be used for RASR
- more reliable modelling of the inter-role transitions