Utterance Verification in Language Learning Applications

Joost van Doremalen, Catia Cucchiarini & Helmer Strik

Department of Linguistics
Radboud University Nijmegen

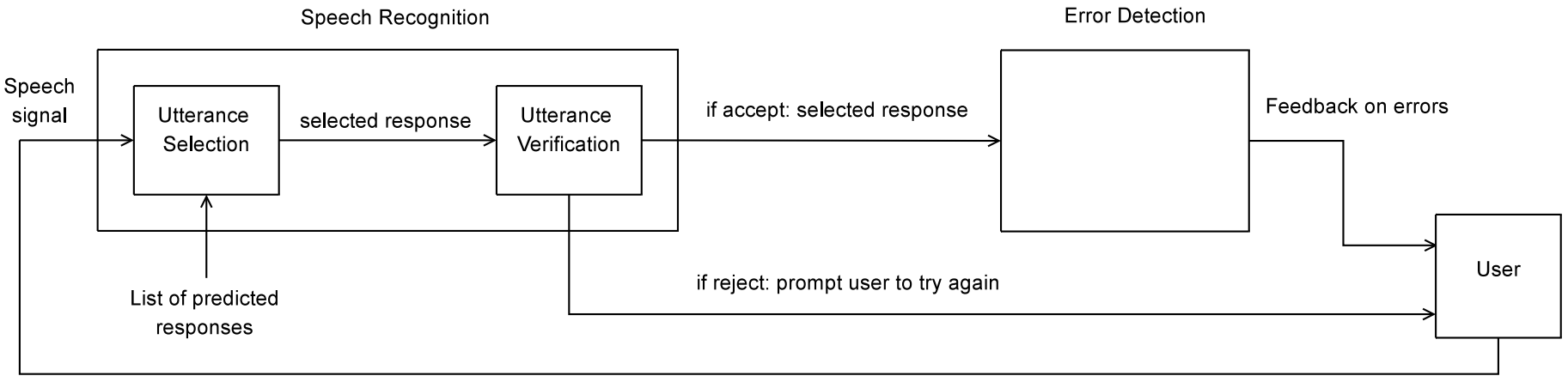
September 1, 2009

Outline

Context: The DISCO project

ASR-based CALL application for training oral proficiency for Dutch as a second language that provides intelligent feedback on pronunciation, morphology and syntax

Context: Experimental System



- **Utterance Selection:** Select response from list of predicted responses
  - *Language Model*: FSM with paths for every response from list
  - *Acoustic Model*: GMMs trained on native read speech and retrained on non-native speech
- **Utterance Verification:** Verify whether selected response reflects what has been said

Context: Why Utterance Verification?

Reject misrecognized utterances:

- When user response not included in list:
  - Response contains a lot of disfluencies, e.g. *repetitions* and *repairs*
  - Sequence of words was not predicted
- When user response included in list but not selected:
  - Phonetically similar to other response(s)

Related work on Utterance Verification

- Accept/reject utterance based on *confidence measure*: number indicating the confidence that the utterance is correctly recognized
- Utterances with confidence measure below certain threshold are rejected
- Several approaches for calculating confidence measures:
  - Combining confidence predictors
    1. hypothesis density
    2. language model related scores
    3. duration information
  - Posterior probability estimation

Related Work: Posterior probability estimation

Speech decoder

$$\hat{w} = \arg\max_{w \in W} [p(w|o)] \tag{1}$$

$$= \arg\max_{w \in W} \left[\frac{p(o|w)p(w)}{p(o)} \right] \tag{2}$$

$$= \arg\max_{w \in W} [p(o|w)p(w)] \tag{3}$$

- How to estimate $p(o)$?
  - Likelihood of optimal phone string using free phone recognizer (*Young*)

Method: Our approach

- Likelihood ratio of response and optimal phone string (LR) as confidence predictor
- Combine using logistic regression:
  - LR
  - Duration information
- How to incorporate duration information:
  - Forced alignment of read speech, 5th and 95th percentile for each phone
  - Number of very short (*nr_shorter_5*) and long (*nr_longer_95*) phones normalized by total number of phonemes

Method: Experimental Setup I

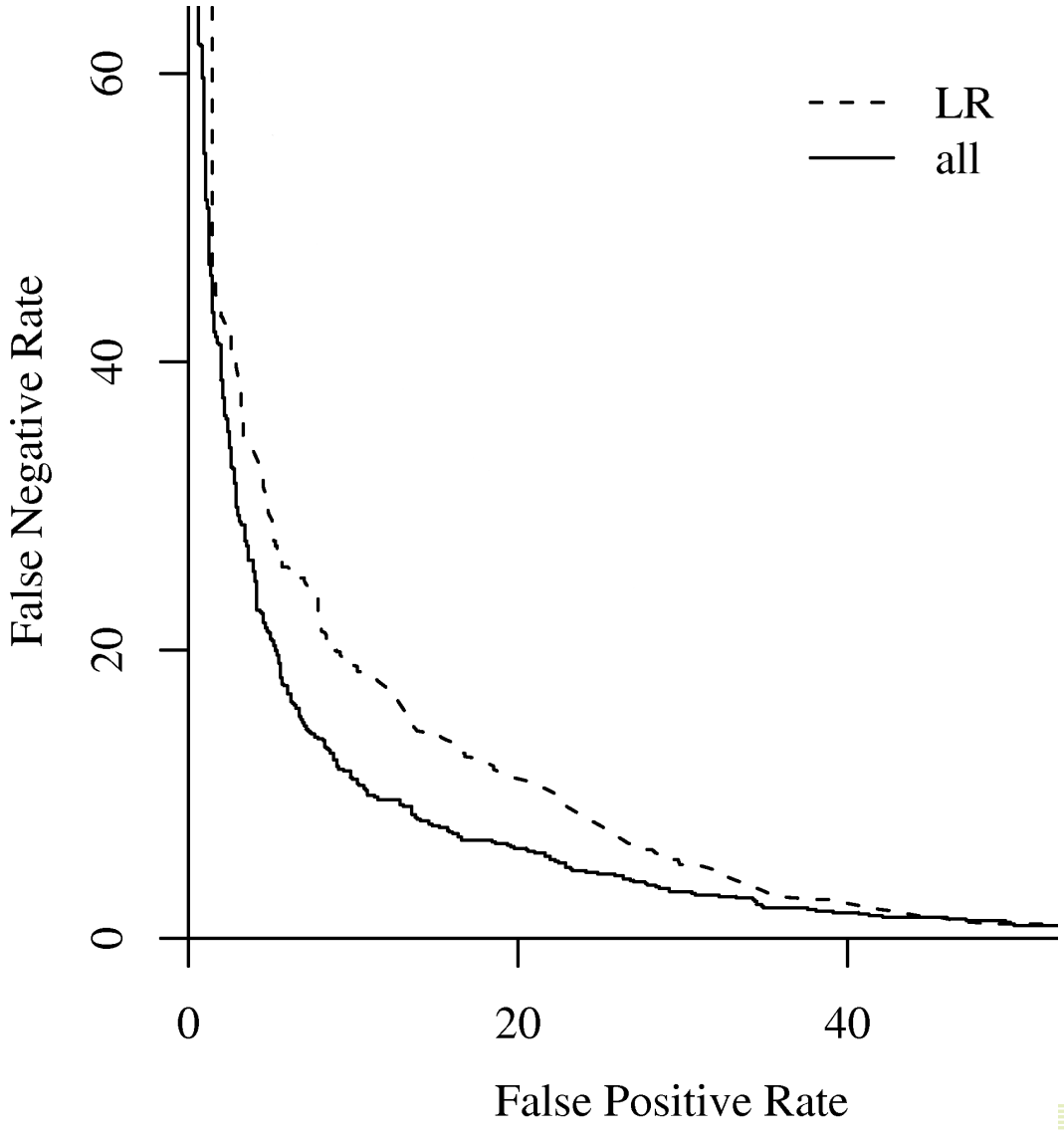
- **Data:**
  - Non-native part of JASMIN corpus
  - 45 speakers giving answers to short questions, 1325 responses
- **Step 1**: Utterance Selection is performed for each response:
  - **with response in language model**: FSM language model (LM) is generated based on the ≈45 responses and task is to select the correct one → succeeds for 90% of utterances
  - **without response in language model**: FSM LM is generated based on the (≈45 responses - the correct reponse)

Method: Experimental Setup II

- **Step 2**: Utterance Verification
  - accept 90% correctly recognized in **with response in LM**
  - reject 10% incorrectly recognized in **with response in LM**
  - reject 100% incorrectly recognized in **without response in LM**
- **Evaluation:** Leave-One-Speaker-Out crossvalidation

Results

Features	EER
LR	14.4%
nr_shorter_5	27.4%
nr_longer_95	35.8%
all	10.3%



Results

- **With response in LM**:

		actual	
		correct	incorrect
decision	accept	80.8%	3.0%
	reject	9.2%	7.0%
		90.0%	10.0%

- **Without response in LM**:

		actual	
		correct	incorrect
decision	accept	-	8.3%
	reject	-	91.7%
		-	100.0%

Discussion

- Other methods to include durations?
  - Mean (absolute) Z-score of utterance
  - Mean probability of duration deviations in utterance

- Threshold calibration:
  - Cost of false accept $C_{FA} \rightarrow$ cost for giving feedback on the wrong utterance?
  - Cost of false reject $C_{FR} \rightarrow$ cost for unnecessarily asking user to try again?
  - Minimize cost function on development set using different thresholds:
    $C_{total} = p_{FR} \cdot C_{FR} \cdot p_{correct} + p_{FA} \cdot C_{FA} \cdot (1 - p_{correct})$

Questions?

Questions?

Formula

- Probability of larger deviation from expected duration:

$$1 - P[\mu_D - |d - \mu_D| < D < \mu_D + |d - \mu_D|]$$

$$1 - \sum_{i=\mu_D-|d-\mu_D|}^{\mu_D+|d-\mu_D|} p_D(i)$$

Language Model Toy Example