BOOSTING OF PROSODIC AND PRONUNCIATION FEATURES TO DETECT MISPRONUNCIATIONS OF NON-NATIVE CHILDREN

Christian Hacker[1], Tobias Cincarek[2], Andreas Maier[1], André Heßler[1], Elmar Nöth[1]

[1]Institute for Pattern Recognition (LME), University of Erlangen-Nuremberg, Germany

[2]Graduate School of Information Science, NAIST, Nara, Ikoma, Japan

Outline

- German children reading English
- Acoustic models: trained on the Pf-Star native corpus (children from Birmingham)
- Pronunciation scoring with **176 features**
- Features from **Prosody**/**Pronfex** Module
- Feature selection with AdaBoost

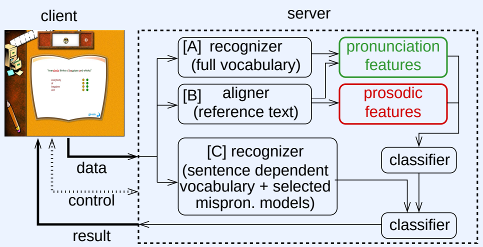

Figure 1: **Caller**: **C**omputer **a**ssisted **l**anguage **le**arning from **Er**langen. Here: Focus on paths via [A] and [B]

Non-Native Data

- 28 children, age 10–11 (English as 2nd lang. since 6 months)
- 72 min; vocabulary: 8088 tokens, 942 types
- reading errors, repetitions, word fragments, non-verbals

10 Raters:

- 1 graduate univ. student of English
- 8 German teachers of English marked words where they would have stopped the student in school
- 1 native British teacher (N)
- **strictness of raters**: 4.4 % – 5.2 %; N: 7.6 % marked as errors

Reference:
Marked if at least 3 experts agree
(5.6 %, cf. strictness)

Class-wise-averaged recognition rate:
CL = 0.5(REC_w + REC_c) (= av. Recall)
c: correctly pronounced; *w*: wrongly pronounced

Agreement:

- rater vs. reference : CL = 77 % (open average)
- pairs of raters: inter-rater: CL = 70 %; intra-rater.: 78 – 80 %
- $REC_w \leq 78\%$; $REC_c \leq 99\%$

AdaBoost

- Select **weak classifiers** $h_t(.)$ that use complementary info
- Here: each $h_t(.)$ is **trained on exactly 1 feature**
- $h_t(x)$ = 1 if mispronounced; 0 else

1. Optimal threshold for each $h_t(.)$ (criterion: CL)
2. A weight $w_{0,i}$ is assigned to each word i of the training data. Weights of either class are distributed uniformly and sum up to 0.5.
3. Choose the weak classifier $h_t(.)$ with lowest error ϵ_t: Words that are wrongly classified contribute with $w_{t,i}$ to the error.
4. Use greater weights for all wrongly classified words:

$$w_{t+1,i} = w_{t,i}\frac{1-\epsilon_t}{\epsilon_t} \quad ; \quad \alpha_t = \log(\frac{1-\epsilon_t}{\epsilon_t}) \qquad (1)$$

5. Normalize the weights; $t = t + 1$; goto 3.
6. Finally, combination to a strong classifier:

$$x \text{ is mispronounced, if } \sum_t \alpha_t h_t(x) \geq \frac{1}{2}\sum_t \alpha_t \qquad (2)$$

- **Leave-one-speaker-out** (loo) evaluation
- Calculate **mean** α_t for each feature over 28 loo-iterations to get this ranking:

best feature →

Forced alignment

Speech recognizer

Phone bigram prob. of recognized phone sequence

Phone confusion c_j = $P(q_j|p_j, M_w)/P(q_j|p_j, M_c)$
q_j recognized phone
p_j phone in reference
$j = 1 \ldots N$: Index of phone

PhoneConf	Mean per word: $1/N \sum_{j=1}^{N} c_j$
Likelihood	Word score
Confidence	Posterior score [-1,0]
ProsEnergy	FFT coefficient 1
Duration	Deviation of phones (scatter)
Confidence	Posterior score of reference word
ProsEnergy	FFT coefficient 0
ProsPitch	Regression of the f_0 [-1,0]
PhoneSeq	Bigram prob. of phones / #phones
ProsPosition	Position of the max. f_0 [1,1]
ProsEnergy	Minimum [-2,-1]
ProsDuration	Normalized [-1,0]
PhoneConf	Maximum per word: $\max_j c_j$
ProsEnergy	Mean [1,2]
PhoneConf	Minimum per word: $\min_j c_j$

Energy: lowest FFT coeff. mean/min in preceding/succ. words

Duration

Pitch/F0 : slope (prec. word); position of max. (succ. word)

Figure 2: Top 15 features selected with AdaBoost

Features Extraction

- **Recognized word chain**, cf. Fig. 1 [A]
  - native models, LME-recognizer, 46 % word acc.
  - 2500 additional mispronunciation models
- **Forced alignment**, cf. Fig. 1 [B]
- **Duration** and **Energy statistics**
  - estimated on native data
- **Phoneme bigram model**
  - estimated on reference texts plus further data
- **Phone confusion statistics** on
  - mispronounced words M_w
  - correctly pronounced words M_c

The Pronfex Module (63 features)

- Rate-of-speech, long pauses
- Duration (deviation from native statistics)
- Log-likelihood of reference word
- Word/phone accuracy and correctness
- Confidence of reference (N-best lists)
- Phone bigram probability
- Phone confusion using M_w and M_c

The Prosody Module (113 features)

- Pitch (F0), energy, duration, jitter, shimmer, pauses (e.g. min, max, mean, regression)
- Position of maxima, minima, on-set, ...
- FFT coefficients of the energy
- Context: preceding word [-1,-1], 2 succ. words [1,2]

Results

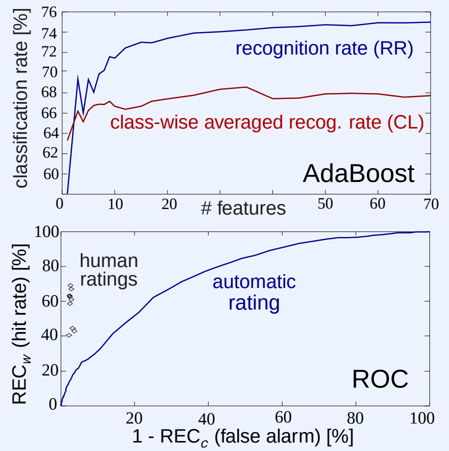

Figure 3: Classification rate with different numbers of features; ROC evaluation with 35 features

- Low CL with prosodic features, but useful extension to pronunciation features
- Best feature: **phone confusion**
- AdaBoost: **No overfitting** to training data
- **Similar feature sets** are selected
  - in all loo-iterations
  - using different references/experts
- 15 features: 66.7 % CL
  35 features: **68.6 % CL**
  → **89 % of human expert agreement**
- Teachers have high agreement on correct words (low hit rate on mispron.'s)