Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer

Stefan Steidl
steidl@informatik.uni-erlangen.de

Chair for Pattern Recognition
University of Erlangen-Nuremberg

Introduction

Problem: Shortage of speech data to train a speech recognizer well

Examples:

- Non-native speech
- Children's speech

Introduction

Problem: Shortage of speech data to train a speech recognizer well

Examples:

- Non-native speech
- Children's speech

Possible solutions:

Introduction

Problem: Shortage of speech data to train a speech recognizer well

Examples:

- Non-native speech
- Children's speech

Possible solutions:

- Collect more data 

Introduction

Problem: Shortage of speech data to train a speech recognizer well

Examples:

- Non-native speech
- Children's speech

Possible solutions:

- Collect more data 
- Add "foreign" data to the training set

Introduction

Problem: Shortage of speech data to train a speech recognizer well

Examples:

- Non-native speech
- Children's speech

Possible solutions:

- Collect more data 
- Add "foreign" data to the training set
- Use adaptation techniques like MLLR, MAP, or VTLN

Introduction

Problem: Shortage of speech data to train a speech recognizer well

Examples:

- Non-native speech
- Children's speech

Possible solutions:

- Collect more data 
- Add "foreign" data to the training set
- Use adaptation techniques like MLLR, MAP, or VTLN
- HMM-Interpolation:
  Interpolate the acoustic models with the models of a second recognizer from a different application scenario

Overview

- Objective
- Combination of Two Codebooks
- Interpolation of the Acoustic Models
- Data
- Experimental Results
- Future Work & Conclusion

Objective

The idea to interpolate the acoustic models is not new:

From literature: K. Livescu (1999), L. Mayfield Tomokiyo (2001)

- Recognition of non-native speech
- One fixed interpolation partner
- Only one single interpolation weight

$$\underbrace{\begin{pmatrix} \theta/\text{ɪ}/\text{ŋ} \\ \text{w}/\text{ɪ}/\text{ŋ} \\ \text{j}/\text{e}/\text{s} \\ \vdots \end{pmatrix}}_{\text{non}-\text{native}} \cdot\lambda + (1-\lambda) \cdot \underbrace{\begin{pmatrix} \theta/\text{ɪ}/\text{ŋ} \\ \text{w}/\text{ɪ}/\text{ŋ} \\ \text{j}/\text{e}/\text{s} \\ \vdots \end{pmatrix}}_{\text{native}}$$

Our new approach:

- An arbitrary number of interpolation partners
- A data driven way of choosing the best interpolation partners
- An automatic method to estimate the interpolation weights

$$\underbrace{\left(\,\theta/\imath/\eta\,\right)}_{\text{non}-\text{native}}\cdot\lambda_1+\lambda_2\cdot\underbrace{\left(\,\theta/\imath/\eta\,\right)}_{\text{native}}+\lambda_3\cdot\underbrace{\left(\,\mathsf{s}/\imath/\eta\,\right)}_{\text{native}}+\lambda_4\cdot\underbrace{\left(\,\mathsf{f}/\imath/\eta\,\right)}_{\text{native}}$$

$$\underbrace{\left(\,\mathsf{w}/\imath/\eta\,\right)}_{\text{non}-\text{native}}\cdot\mu_1+\mu_2\cdot\underbrace{\left(\,\mathsf{w}/\imath/\eta\,\right)}_{\text{native}}+\mu_3\cdot\underbrace{\left(\,\mathsf{v}/\imath/\eta\,\right)}_{\text{native}}$$

Problems

1. Two distinct speech recognizers with semi-continuous HMMs:

   $\Rightarrow$ Two codebooks with 500 probability density functions each

   $\Rightarrow$ Combination of both codebooks

2. Estimation of the interpolation weights

3. Choice of the interpolation partners

Combination of Two Codebooks

- 1:1-mapping of the two codebooks C_1 and C_2

Combination of Two Codebooks

- 1:1-mapping of the two codebooks C_1 and C_2
- Greedy algorithm:
  - Look for the "best" matching pair $(\mathcal{N}_1, \mathcal{N}_2)$ of the remaining density functions, $\mathcal{N}_1 \in C_1$, $\mathcal{N}_2 \in C_2$
  - Combine $\mathcal{N}_1$ and $\mathcal{N}_2$ by averaging their parameters
  - Repeat until all densities are matched

Combination of Two Codebooks

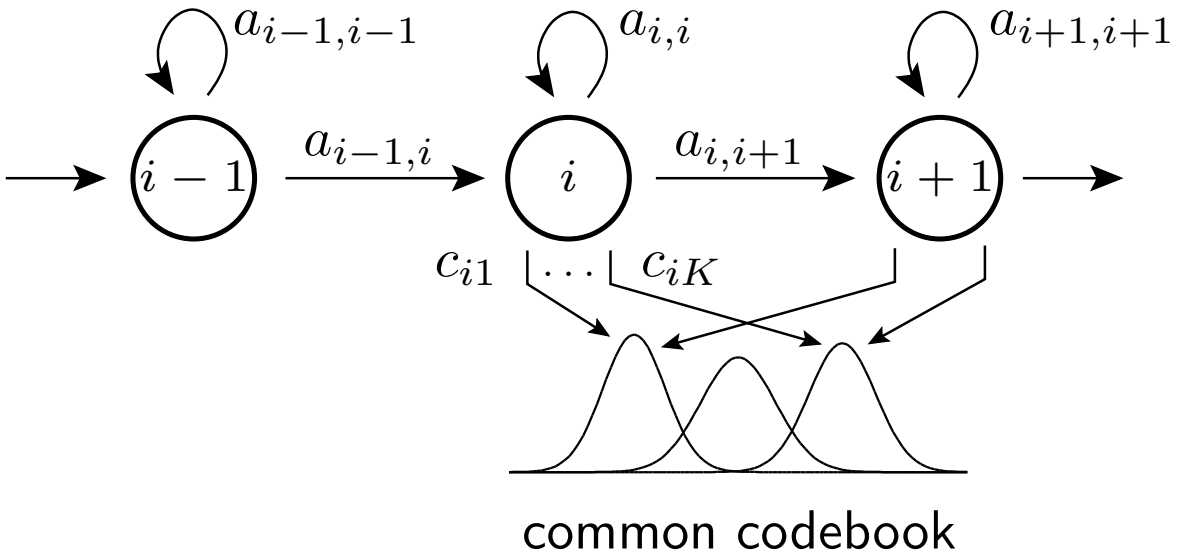
- 1:1-mapping of the two codebooks C_1 and C_2
- Greedy algorithm:
  - Look for the "best" matching pair $(\mathcal{N}_1, \mathcal{N}_2)$ of the remaining density functions, $\mathcal{N}_1 \in C_1$, $\mathcal{N}_2 \in C_2$
  - Combine $\mathcal{N}_1$ and $\mathcal{N}_2$ by averaging their parameters
  - Repeat until all densities are matched
- Distance measure: Increase of entropy

Combination of Two Codebooks

- 1:1-mapping of the two codebooks C_1 and C_2
- Greedy algorithm:
  - Look for the "best" matching pair $(\mathcal{N}_1, \mathcal{N}_2)$ of the remaining density functions, $\mathcal{N}_1 \in C_1$, $\mathcal{N}_2 \in C_2$
  - Combine $\mathcal{N}_1$ and $\mathcal{N}_2$ by averaging their parameters
  - Repeat until all densities are matched
- Distance measure: Increase of entropy
- Option: Different weighting of both codebooks (codebook interpolation)

Interpolation of the Acoustic Models (1)

- Semi-continuous hidden Markov models



Mixture weights c_{ik} and transition probabilities a_{ij}

Interpolation of the Acoustic Models (2)

- Interpolate state s_i with the corresponding states s_{i_j} of the $J-1$ partners

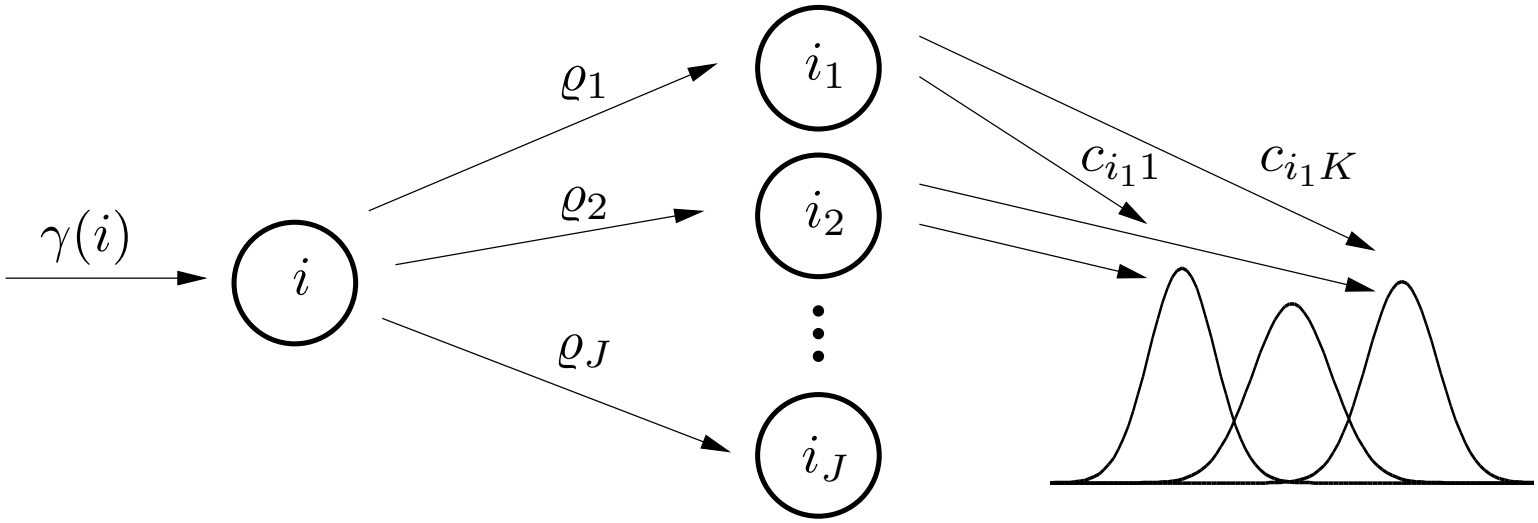
  1st step: Interpolate the mixture weights c_{ik}

$$\hat{c}_{ik} = \varrho_1 \cdot c_{ik} + \varrho_2 \cdot c_{i_2 k} + \ldots + \varrho_J \cdot c_{i_J k} \quad \text{with} \quad \sum_{j=1}^{J} \varrho_j = 1$$

  EM algorithm: Automatic estimation of the weights ϱ_j on a validation set

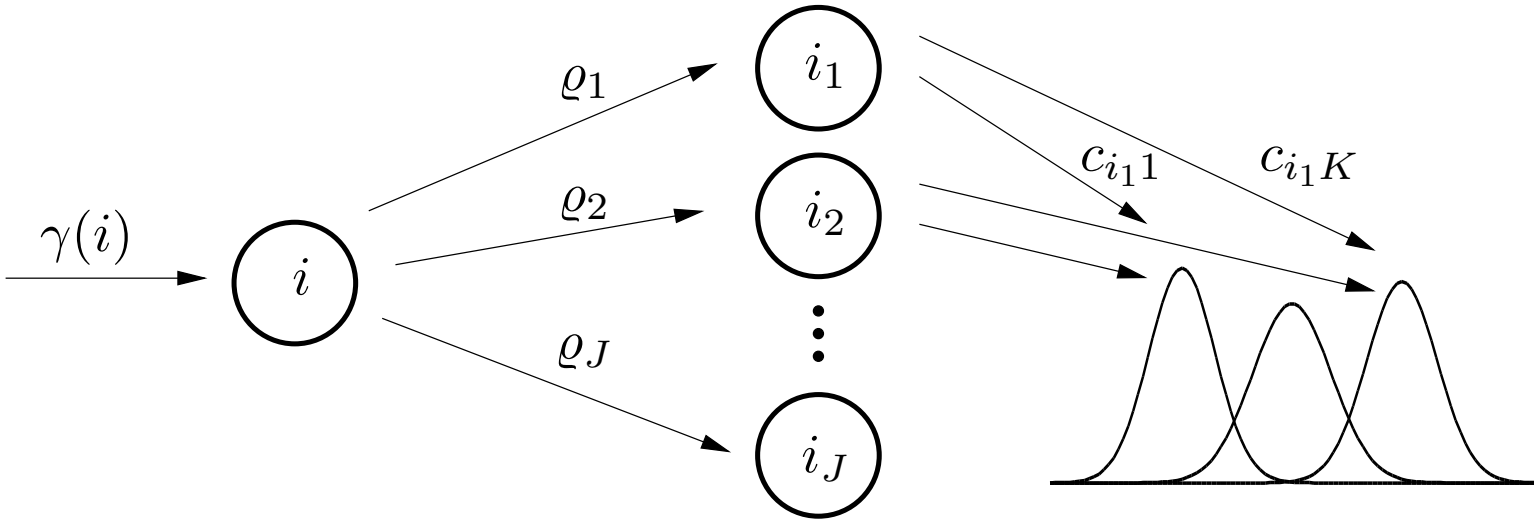
  2nd step: Interpolate the transition probabilities a_{ij}

EM Algorithm



$$\varrho_j \quad = \quad P(s_{i_j} \mid s_i, \boldsymbol{\varrho}) = \sum_{k=1}^{K} P(k \mid s_i, \boldsymbol{\varrho}) \cdot \frac{P(s_{i_j}, k \mid s_i, \boldsymbol{\varrho})}{P(k \mid s_i, \boldsymbol{\varrho})}$$

EM Algorithm



$$\varrho_j = P(s_{i_j} \mid s_i, \boldsymbol{\varrho}) = \sum_{k=1}^{K} P(k \mid s_i, \boldsymbol{\varrho}) \cdot \frac{P(s_{i_j}, k \mid s_i, \boldsymbol{\varrho})}{P(k \mid s_i, \boldsymbol{\varrho})}$$

$$\tilde{\varrho}_j = \sum_{k=1}^{K} \zeta(i,k) \cdot \frac{\varrho_j \cdot c_{i_j k}}{\sum_{j=1}^{J} \varrho_j \cdot c_{i_j k}}$$

Evaluation of the HMM Interpolation

Quality function:

$$\ell(\varrho_1, \ldots, \varrho_J) \quad = \quad \log \prod_{k=1}^{K} \Bigg(\underbrace{\sum_{j=1}^{J} \varrho_j \cdot c_{i_j k}}_{c_{ik}} \Bigg)^{\zeta(i,k)}$$

The quality is high if the distributions c_{ik} and $\zeta(i,k)$ are similiar

Advantage: No expensive calculation of $P(\boldsymbol{X}|\boldsymbol{\lambda})$ on the validation set

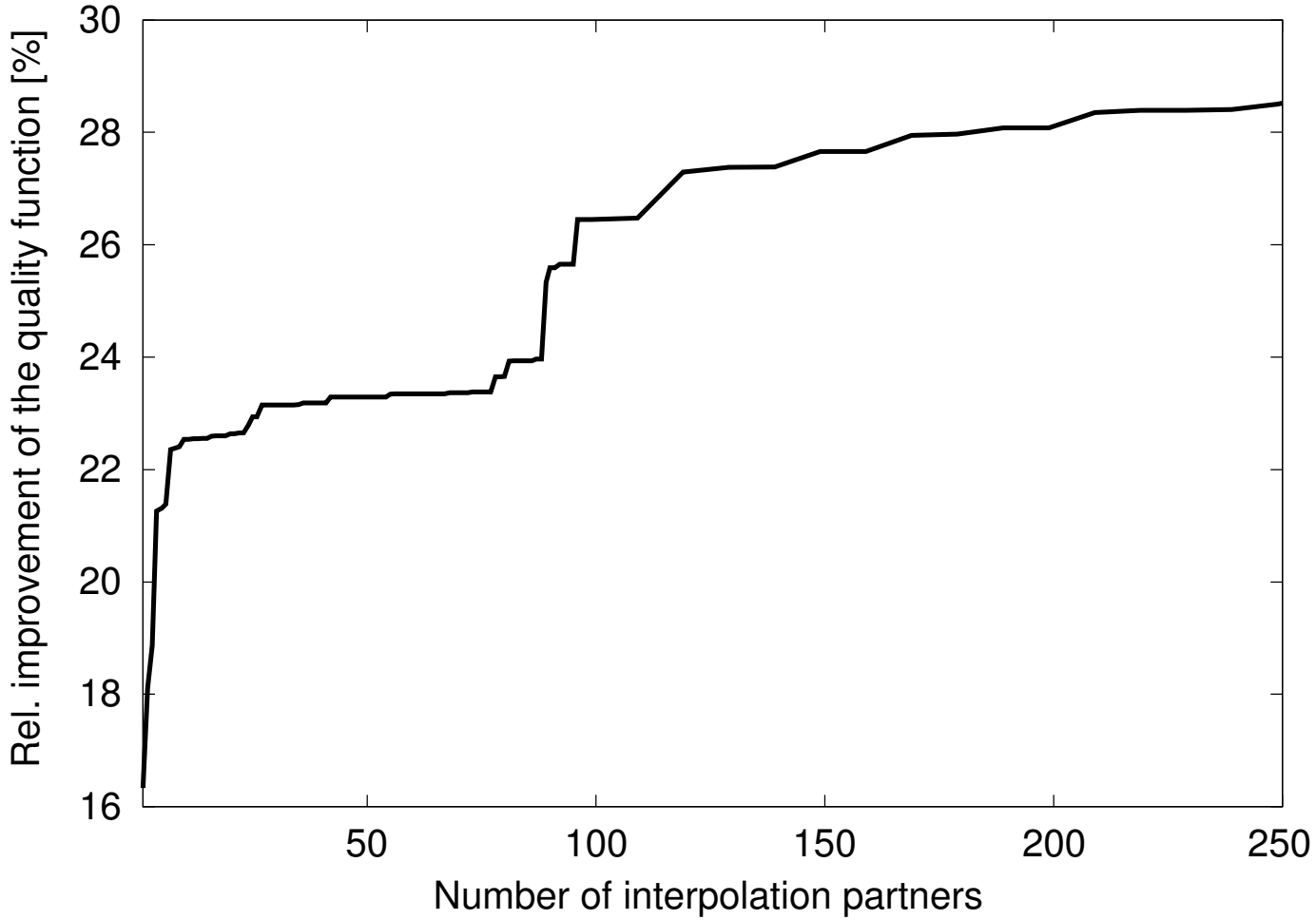
Choice of the Interpolation Partners (1)

- Interpolate each children's polyphone (567) with each adults' polyphone (11,954) and evaluate the quality function
- Create a n-best list
- Example: n-best list for monophone o:

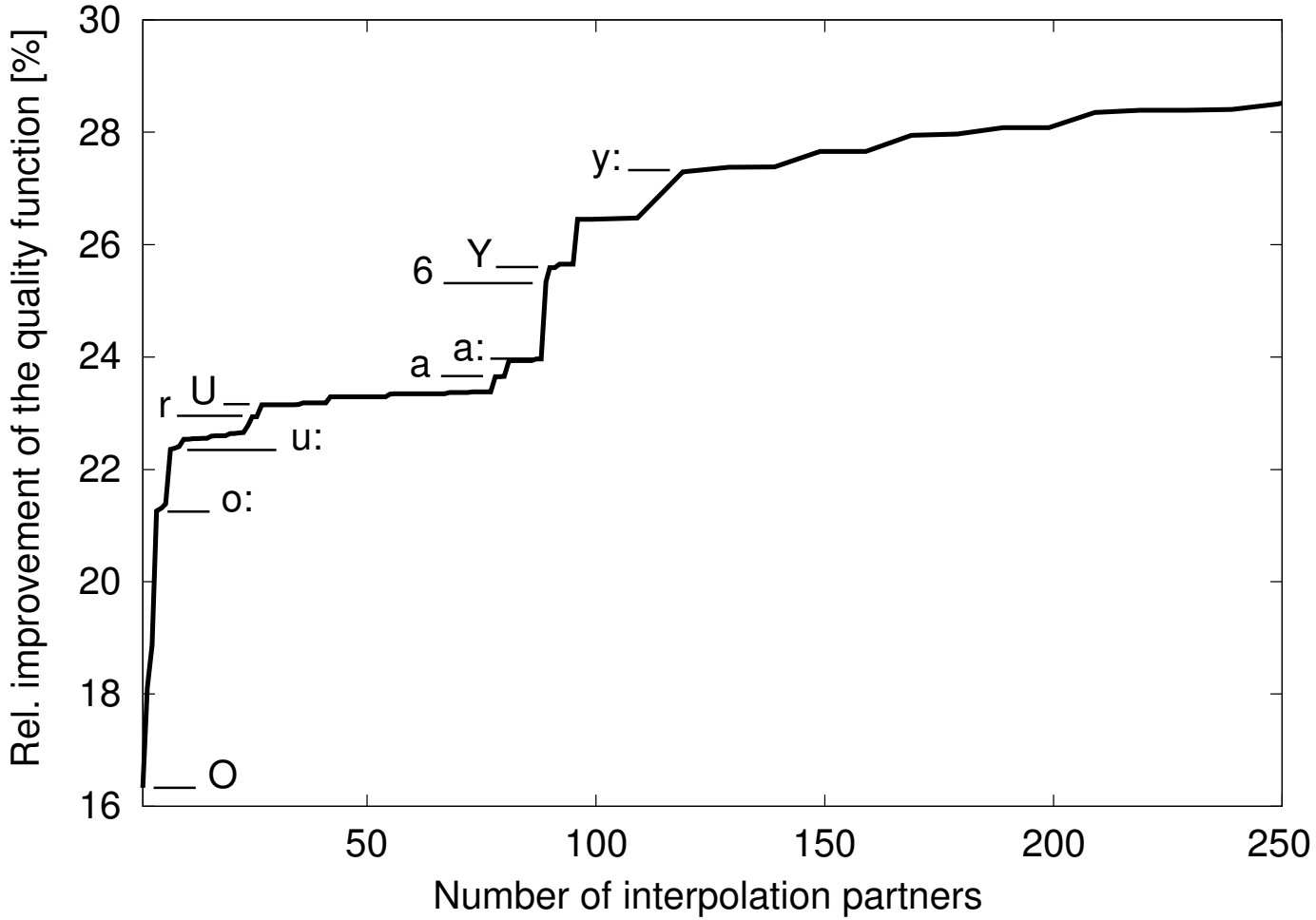
1	/O/K	16.3 %
2	|d/O/x	16.2 %
3	K/O/k	14.9 %
4	z/o:/|K	14.8 %
5	|z/o:/|v	14.7 %
6	/O/k	14.3 %
7	T/u:/k	14.2 %

25	V|/r/Um	11.3 %
27	z/U/N	11.1 %
78	#d/a/|fo6|	8.7 %
87	#d/a:/|fF	7.6 %
89	b/6/|K	7.0 %
90	v/Y/rK	6.9 %
119	r/y:/|K	6.0 %

Choice of the Interpolation Partners (2)



Choice of the Interpolation Partners (2)



Choice of the Interpolation Partners (3)

Observations:

- Restriction to ≤ 50 interpolation partners is not optimal
- Noticeable steps caused by polyphones with identical core phone and similar context

z/**o:**/|K	14.8 %
|z/**o:**/|v	14.7 %
both	15.9

Choice of the Interpolation Partners (3)

Observations:

- Restriction to ≤ 50 interpolation partners is not optimal
- Noticeable steps caused by polyphones with identical core phone and similar context

z/**o:**/|K	14.8 %
|z/**o:**/|v	14.7 %
both	15.9

Solution:

- Thin out the n-best list:
  Accept only those polyphones with a certain minimal distance to the polyphones already chosen
- Distance measure: Kullback-Leibler divergence

Data (1)

Objective: Interpolation of our children's speech recognizer with the Verbmobil recognizer for adults' speech

Adults' speech:

- VERBMOBIL: speech-to-speech translation project
- Data: 28 hours of spontaneous speech

Children's speech:

- 62 children read *Nordwind & Sonne* and three texts of the *Zürcher Lesetests*
- Vocabulary: 227 words
- Data: 3.5 hours of read speech

Data (2)

Partitioning of the data

Task	Speakers	Text
Training of the children's speech recognizer	40 6 of 40	*Zürcher Lesetest* *Nordwind & Sonne*
HMM interpolation	6 - 40	*Nordwind & Sonne*
Evaluation	20	*Nordwind & Sonne*

Baseline system: Initialization with the Verbmobil codebook

- Vocabulary reduced to *Nordwind & Sonne* (74 words)
- Zerogram language model
- 74.6 % word accuracy on the evaluation set

Experimental Results (1)

HMM interpolation: Influence of the number of interpolation partners

Experiment	Word accuracy
Baseline	74.6 %
1 Partner	79.2 %
5 Partner	79.8 %
10 Partner	80.1 %
20 Partner	**80.9 %**
30 Partner	80.7 %
40 Partner	**80.9 %**
50 Partner	80.8 %

Experimental Results (2)

HMM interpolation: Influence of the size of the validation set (using 20 interpolation partners)

Validation set	Word accuracy
6 speakers	**81.4 %**
12 speakers	81.1 %
18 speakers	80.7 %
24 speakers	81.0 %
30 speakers	80.8 %
40 speakers	80.9 %

Future Work

- Comparison with
  - "simple" HMM interpolation (1 fixed interpolation partner, only 1 interpolation weight)
  - other adaptation methods like MAP, MLLR, and VTLN
- Combination of HMM interpolation and MLLR/VTLN
- Separate HMM interpolation for different groups of speakers
- Repetition of our experiments using our new children corpus with a larger vocabulary

Conclusion

- Precondition: Mapping of both codebooks
- HMM interpolation:
  - Arbitrary number of interpolation partners
  - Data-driven method to find the best interpolation partners
  - EM algorithm to estimate the interpolation weights
- Relative improvement of the word accuracy of 9.1 %

The End