T-61.182 Robustness in Language and Speech Processing

Daniel Tapias Merino

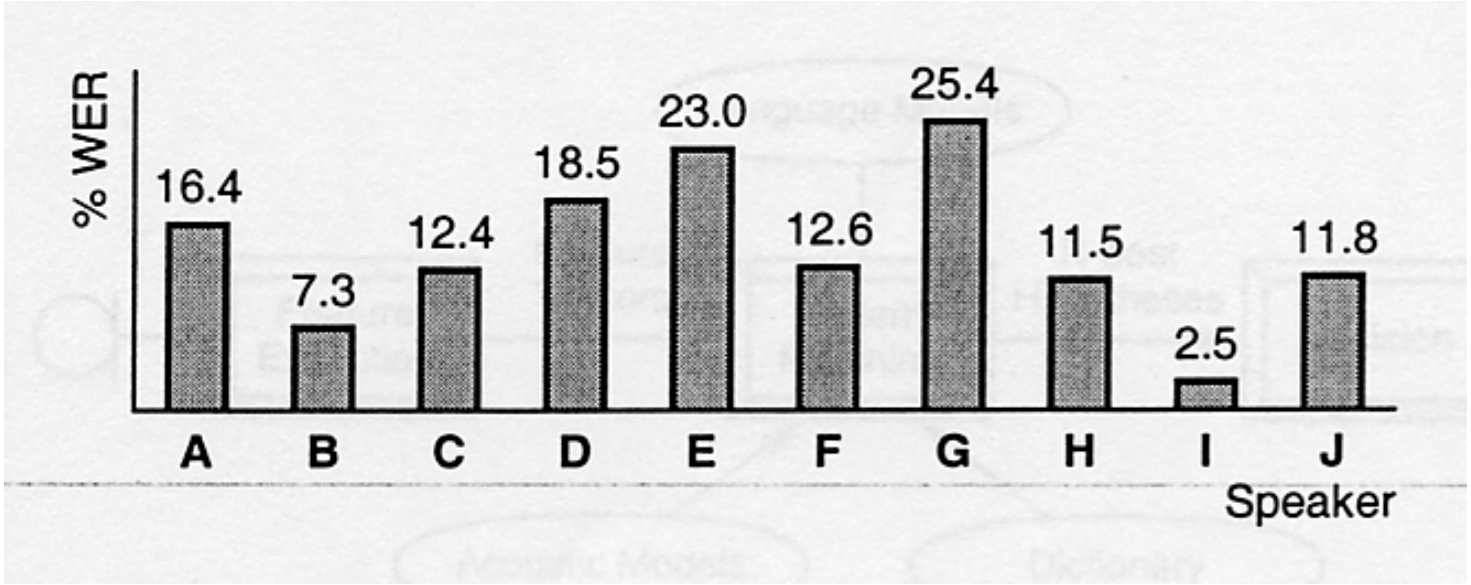
Speaker Compensation in Automatic Speech Recognition, December 2002

explained by

Ramūnas Girdziušas

February 20th, 2003

1 – Introduction



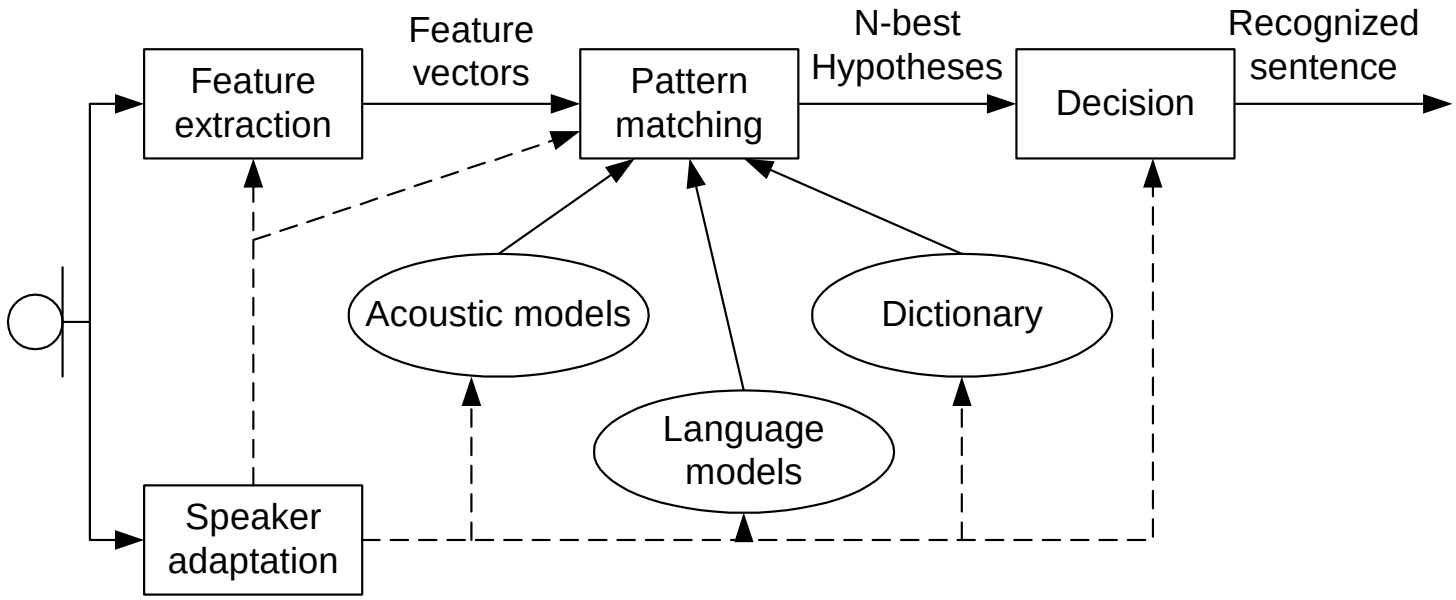
Kuva 1: Speaker independent ASR performs very badly on some speakers. How to compensate inter-speaker variability in ASR?

2 – General things about speaker compensation

2.1 – Causes of inter/intra-speaker variation

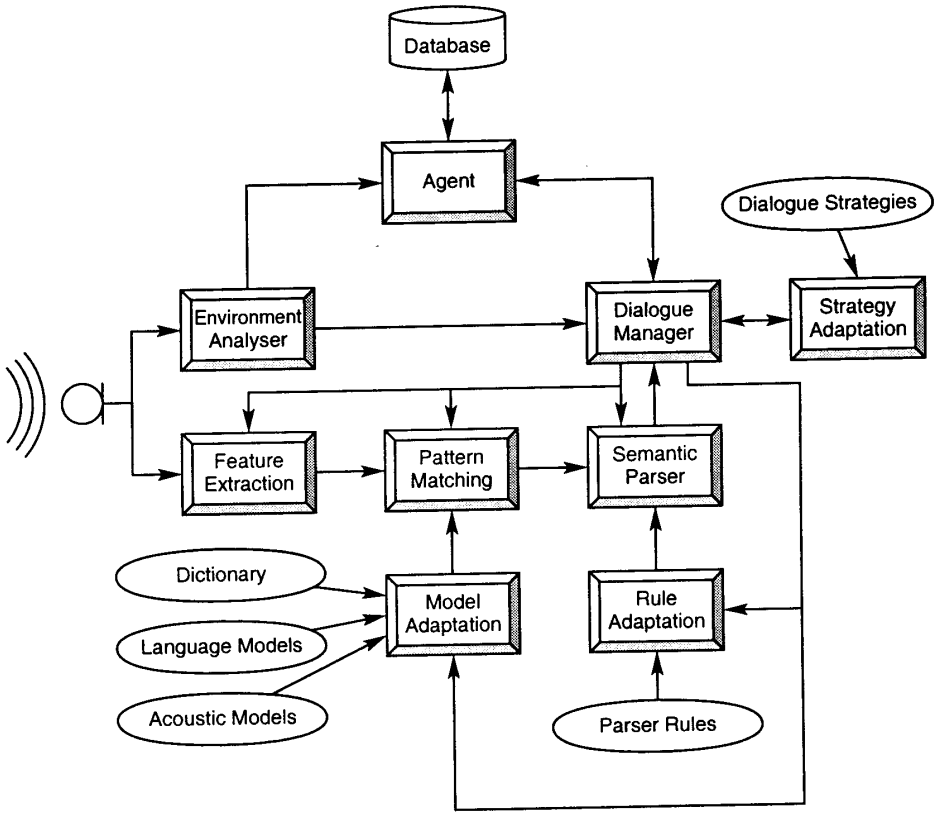
- **Cultural differences**
  - speech loudness, rate, intonation.
  - set of sounds, their duration.
  - way of building sentences.
  - phonetic phenomena such as sound deletion, cl. memory → memry
- **Physiological differences**
  - vocal tract shape, length.
  - age group, tiredness.
- **Environmental differences**
  - background noise, room acoustics, 'cognitive load'.

2.2 – Speaker-adaptive ASR



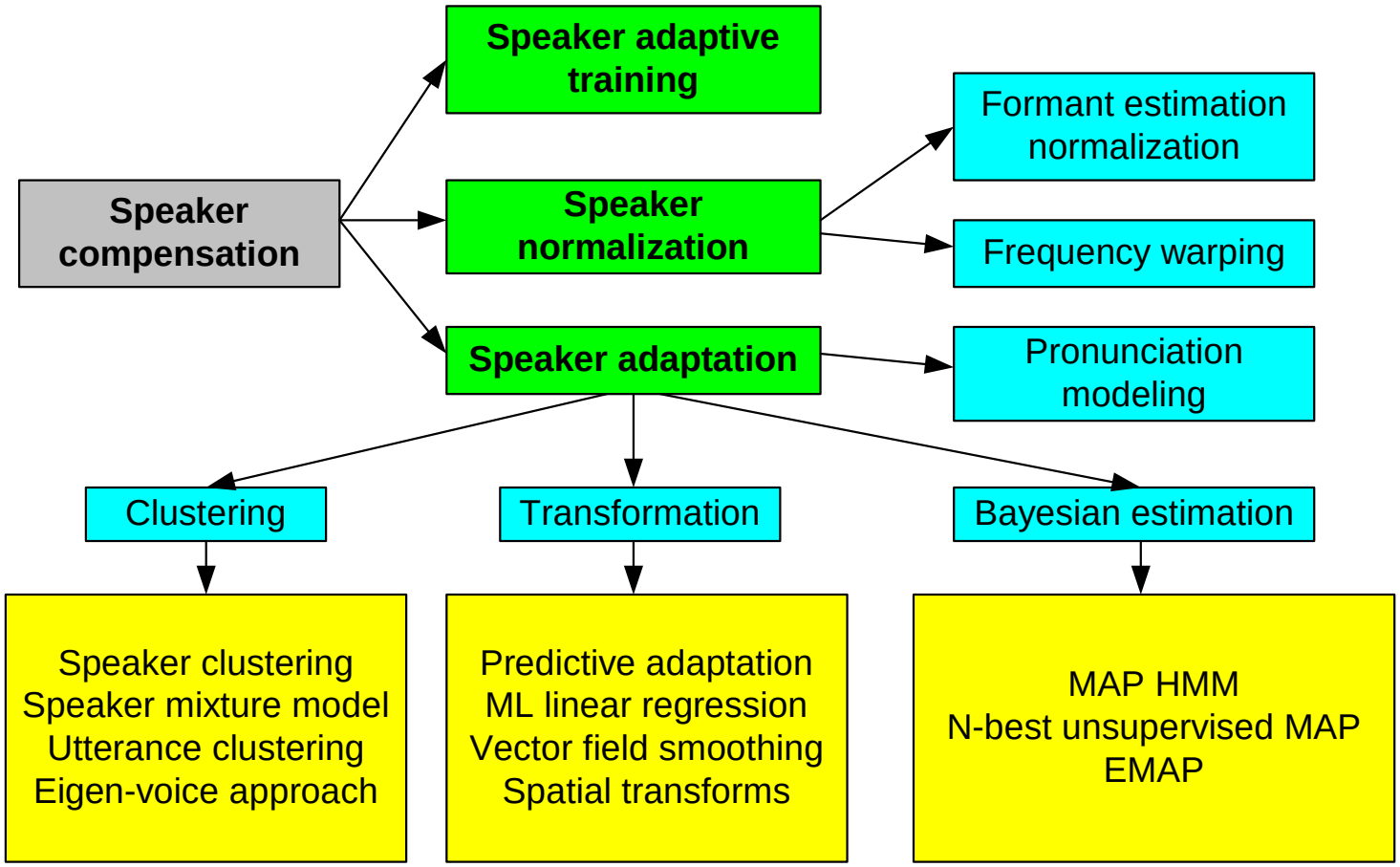
Kuva 2: Most of the adaptation techniques are applied to feature extraction and acoustic models.

2.3 – More general speaker-adaptive ASR



Kuva 3: Block diagram of a future ASR.

2.4 – Speaker compensation techniques



Kuva 4: Bayesian estimation will be skipped entirely.

3 – Speaker adaptation: clustering

3.1 – Speaker clustering

Look for the speaker cluster closest to the new voice:

1. Compute speaker-dependent models.
2. Decode the adaptation data with speaker independent ASR.
3. Viterbi align the adaptation data against the transcriptions.
4. Compute the acoustic likelihood of the adaptation data conditioned on the alignment for each training speaker.
5. Rank the training speakers, choose top N speakers.

Update cluster members by using the MLLR technique.

Test results

- Vocabulary: 20000 words, 35000 utterances, 284 speakers
- 50 training speakers, 3 adaptation utterances.
- 19.5% relative improvement with speaker-independent models.
- 30% relative improvement with speaker-dependent models.

Similar to speaker clustering is a Speaker Mixture model

- Linear combination of speaker dependent models:

$$b_j(\mathbf{x}_t) = \sum_s w_j^s b_j^s(\mathbf{x}_t). \quad (1)$$

- Weights are retrained using data of a new speaker.
- Low-weight training speaker models are removed.

3.2 – Other clustering approaches

Condition dependent utterance clustering

- Creation of condition-dependent models.
- Train stress detector $p(\epsilon_k|w_i)$ for each word w_i.
- Recognize speech by using a stress-dependent recognizer.
- 34% relative improvement on TIMIT+SUSAS, 35 words, 1 isolated word in adaptation.

Eigen-voices approach

- Train R speakers, get their R D-dimensional parameter vectors.
- Apply PCA to get R D-dimensional eigen-voices.
- Represent new speaker as a linear combination of eigen-voices.
- 26.7% average improvement, 120 training speakers and 30 testing speakers, 4-letter adaptation utterances, 6 eigen-voices.

4 – Speaker adaptation: transformation methods

4.1 – Maximum likelihood linear regression

1. Train speaker-independent CDHMM.
2. Transform the model parameters $\mu_s = W_s \epsilon_s$ in order to maximize the model likelihood on the adaptation data, s - number of regression classes.

Test results

- Vocabulary: 1000 words, 10 regression classes.
- 3990 training utterances, 40 adaptation utterances.
- 37% improvement by using a speaker-independent ASR.
- 58% improvement with a speaker-dependent ASR.

4.2 – Other transformation-based approaches

Predictive Speaker Adaptation,
Transfer Vector Field Smoothing

Address the problem of sparse adaptation data. Poorly represented or unseen sounds of the new speaker can be predicted by building additional regression models for the mean values of the Gaussian state observation probabilities.

Spatial Relation-based Transformation

The idea is to adapt Gaussian means also in the phoneme context direction based on the additional spatial relation between the context dependent and context independent CDHMMS.

5 – Speaker adaptation: pronunciation modelling

- During the speech alignment, each possible pronunciation is considered, such as 'comprado' → 'comprao'.
- Aligned corpus is used to estimate $p(r)$ for each phone transformation rule r, such as /ado/ → /ao/.
- A decision tree is built for each rule to predict $p(r|w,m)$, w - word, m - speaking mode.
- Probability for the alternative pronunciation $q_i(w)$ is then computed as

$$P(q_i(w)) = \frac{1}{Z} \prod_{\forall r^+} p(r^+) \prod_{\forall r^-} [1 - p(r^-)], \qquad (2)$$

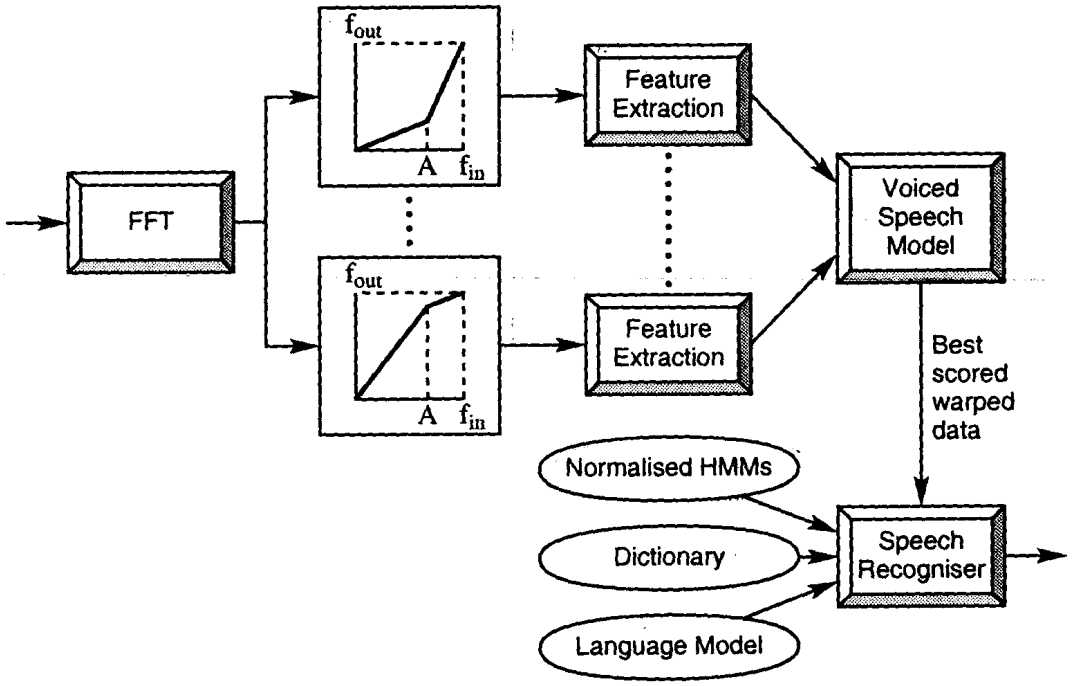
  where r^+ - rules used to obtain $q_i(w)$, r^- - rules that only match with the baseform, w.
- $P(q_i(w))$ is used as a weight during the recognition process.

Test results

- The expanded dictionary included 1.78 alternative pronunciations per word on average.
- Polyphonic decision tree based on the speaking mode and the FTA labels (?) gave best performance.
- CallHome data: WER reduced from 43.7% to 36.1%.
- Switchboard data - from 32.6% to 26.7%.

6 – Speaker normalization techniques

Frequency warping linear transformation



Kuva 5: Frequency warping reduces the effect of the length of the vocal tract. Voiced speech model is trained to select best warp.

Test results

- Switchboard CAIP set data, vocabulary: 10000 words.
- 65 hours of training speech, 80 female and 80 male.
- WER reduction from 49.8% to 43.9%.

Formant estimation normalization

- Scales the frequency axis using the first and second formant frequencies of the new speaker.
- Achieved 3% relative WER reduction on TIMIT database (phone recognition, 426 speakers).

7 – Speaker Adaptive Training

The idea is to reduce the overlap of the acoustic models due to the inter-speaker variation in the *speaker-independent ASR*.

Test results

- Training data: 62 hours of speech from 284 speakers, Wall Street Corpus
- Testing data: 1994 H1 and 1994 S0, 20000 and 5000 words, 20 speakers, 40 adaptation utterances.
- 19% and 26% relative WER reductions achieved.

8 – Conclusions

- **Speaker adaptation** (Clustering, MLLR,...)
  - Modifies parameters of the speaker-independent ASR.
  - Provides large error reductions (30%-90%).
  - Problem: 'real time' constraint.
- **Speaker compensation** (Frequency warping,...)
  - Does not require modifications in speaker-independent ASR.
  - Better suitable for real time applications.
  - Problem: small WER reductions (3%-10%).
- **Speaker adaptive training**
  - Applies speaker compensation techniques to improve WER of the speaker-independent ASR.
  - Still a long way towards a true speaker-independence...