Optimisation of acoustic models for a target accent using decision-tree state clustering

PRASA 2012

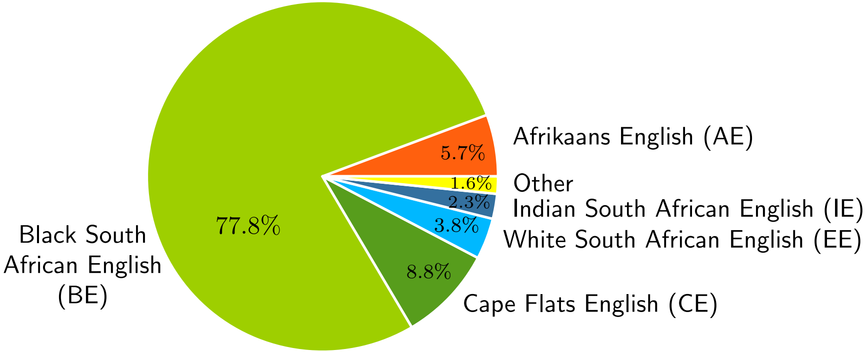
Herman Kamper and Thomas Niesler

Digital Signal Processing Group
Department of Electrical and Electronic Engineering
Stellenbosch University

UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Introduction

Five major **accents of South African English**:



Modelling accents

- How can we **model** the different accents for speech recognition?
- **AST databases**: approximately 6 hours of speech in each accent
- **Multi-accent acoustic modelling** allows selective sharing across accents
- This approach guarantees **overall** likelihood improvement over all accents, but not **per-accent** improvements
- How do we obtain best acoustic model set for **particular accent**, but still incorporate useful data from other accents?

Modelling accents

- How can we **model** the different accents for speech recognition?
- **AST databases**: approximately 6 hours of speech in each accent
- **Multi-accent acoustic modelling** allows selective sharing across accents
- This approach guarantees **overall** likelihood improvement over all accents, but not **per-accent** improvements
- How do we obtain best acoustic model set for **particular accent**, but still incorporate useful data from other accents?

Modelling accents

- How can we **model** the different accents for speech recognition?
- **AST databases**: approximately 6 hours of speech in each accent
- **Multi-accent acoustic modelling** allows selective sharing across accents
- This approach guarantees **overall** likelihood improvement over all accents, but not **per-accent** improvements
- How do we obtain best acoustic model set for **particular accent**, but still incorporate useful data from other accents?

Modelling accents

- How can we **model** the different accents for speech recognition?
- **AST databases**: approximately 6 hours of speech in each accent
- **Multi-accent acoustic modelling** allows selective sharing across accents
- This approach guarantees **overall** likelihood improvement over all accents, but not **per-accent** improvements
- How do we obtain best acoustic model set for **particular accent**, but still incorporate useful data from other accents?

Modelling accents

- How can we **model** the different accents for speech recognition?
- **AST databases**: approximately 6 hours of speech in each accent
- **Multi-accent acoustic modelling** allows selective sharing across accents
- This approach guarantees **overall** likelihood improvement over all accents, but not **per-accent** improvements
- How do we obtain best acoustic model set for **particular accent**, but still incorporate useful data from other accents?

Acoustic modelling

Acoustic modelling of context-dependent phones

- Use hidden Markov models (HMMs)
- Acoustic modelling of **triphones**: [t]−[iy]+[n]
- Problems:
  - Not all triphones occur in the training data
  - Not enough data for some triphones which do occur
- Want to **determine clusters** of similar triphones

Acoustic modelling

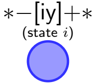
Acoustic modelling of context-dependent phones

- Use hidden Markov models (HMMs)
- Acoustic modelling of **triphones**: [t]−[iy]+[n]
- Problems:
  - Not all triphones occur in the training data
  - Not enough data for some triphones which do occur
- Want to **determine clusters** of similar triphones

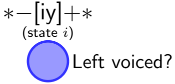
Solution

Use **decision-tree state clustering**

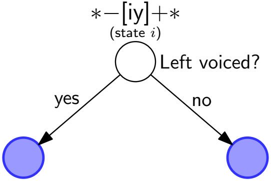
Decision-tree state clustering



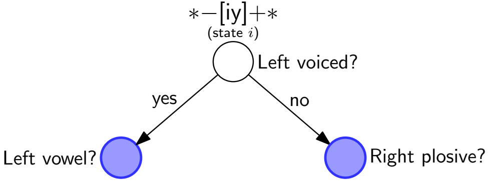
Decision-tree state clustering



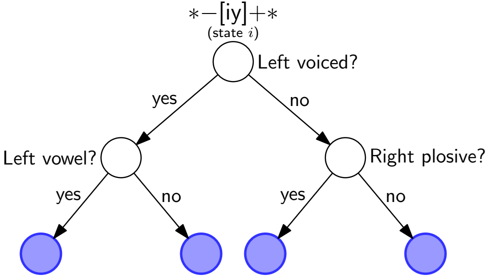
Decision-tree state clustering



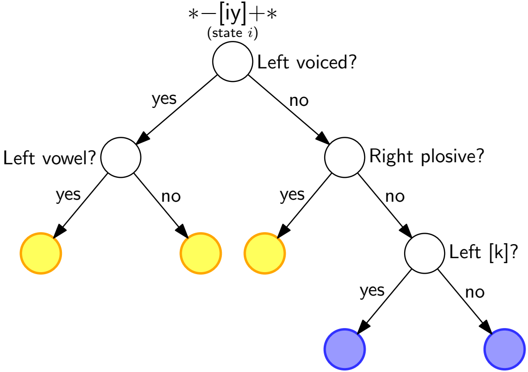
Decision-tree state clustering



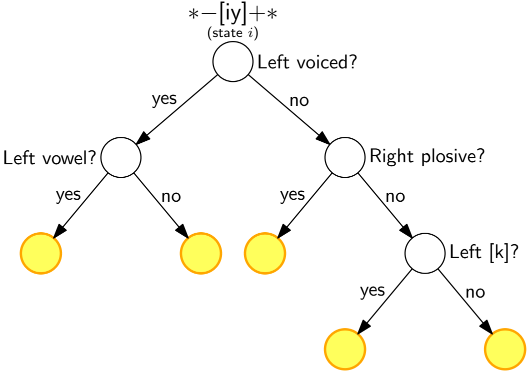
Decision-tree state clustering



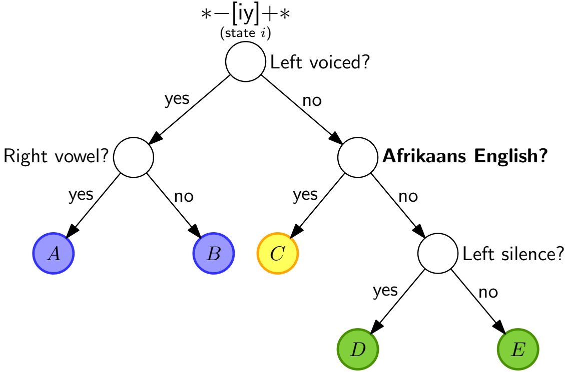
Decision-tree state clustering



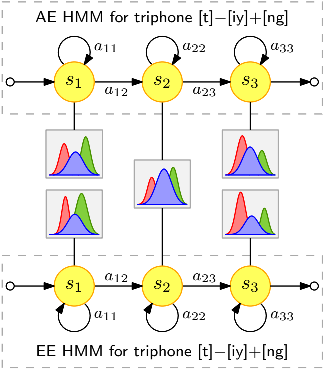
Decision-tree state clustering



Decision-tree state clustering

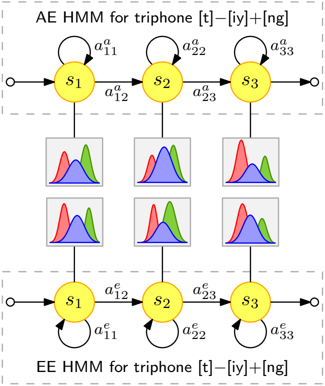


Multi-accent acoustic modelling



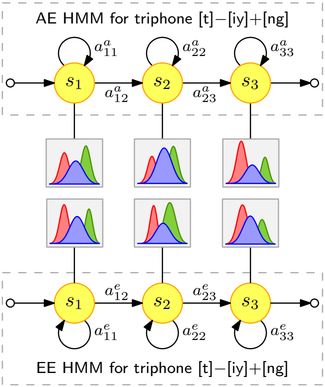
Traditional modelling approaches

Accent-specific models

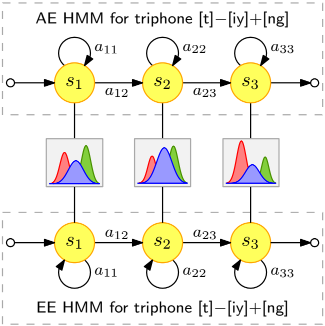


Traditional modelling approaches

Accent-specific models



Accent-independent models



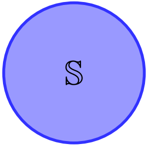
Traditional modelling approaches

Phone recognition accuracy (%)

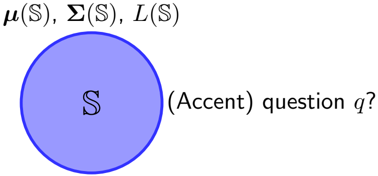
Approach	AE	BE	CE	EE	IE	Average
Accent-specific	64.80	56.77	64.59	72.97	64.27	64.68
Accent-independent	65.97	55.98	66.51	74.45	64.40	65.44
Multi-accent	66.20	56.56	66.31	73.94	64.60	**65.50**

Problem with multi-accent state clustering

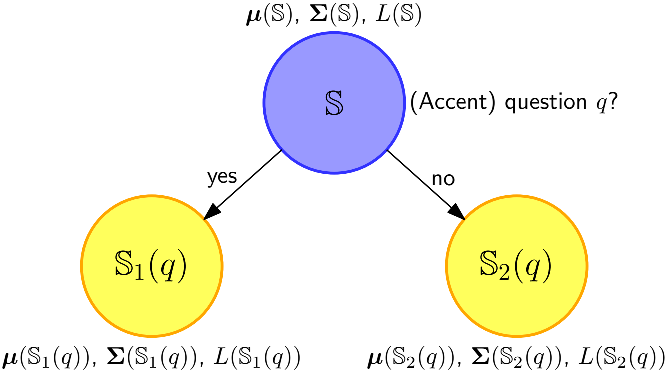
$\boldsymbol{\mu}(\mathbb{S})$, $\boldsymbol{\Sigma}(\mathbb{S})$, $L(\mathbb{S})$



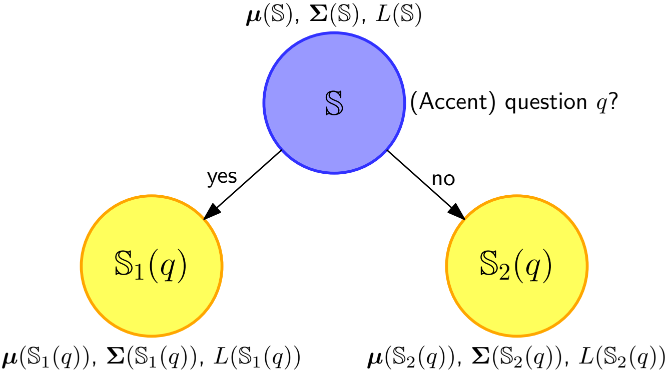
Problem with multi-accent state clustering



Problem with multi-accent state clustering

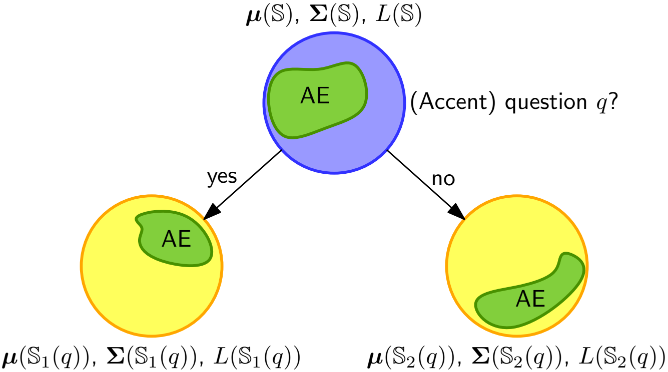


Problem with multi-accent state clustering



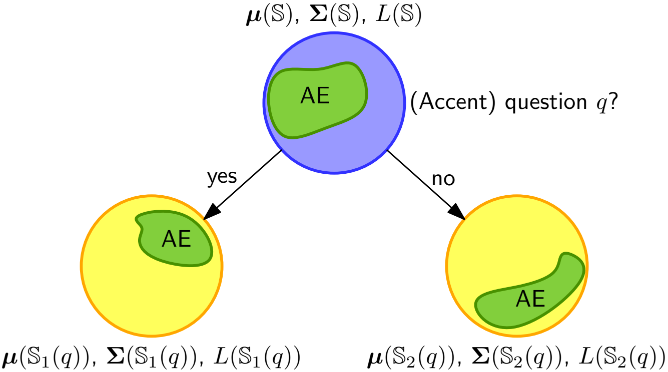
Splitting criterion: $\Delta L_q = L(\mathbb{S}_1(q)) + L(\mathbb{S}_2(q)) - L(\mathbb{S})$

Problem with multi-accent state clustering



Splitting criterion: $\Delta L_q = L(\mathbb{S}_1(q)) + L(\mathbb{S}_2(q)) - L(\mathbb{S})$

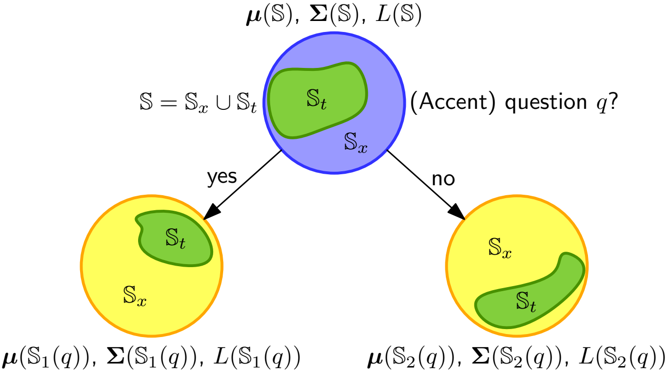
Problem with multi-accent state clustering



Splitting criterion: $\Delta L_q = L(\mathbb{S}_1(q)) + L(\mathbb{S}_2(q)) - L(\mathbb{S})$

The **question** is: what happens to $L_{\text{AE}}(\mathbb{S})$?

Problem with multi-accent state clustering



Splitting criterion: $\Delta L_q = L(\mathbb{S}_1(q)) + L(\mathbb{S}_2(q)) - L(\mathbb{S})$

The **question** is: what happens to $L_t(\mathbb{S})$?

Targeted multi-accent acoustic modelling

Proposal: replace $L(\mathbb{S})$ with $L_t(\mathbb{S})$ in the standard clustering procedure

Targeted multi-accent acoustic modelling

Proposal: replace $L(\mathbb{S})$ with $L_t(\mathbb{S})$ in the standard clustering procedure

But **can we calculate** $L_t(\mathbb{S})$?

Targeted multi-accent acoustic modelling

Proposal: replace $L(\mathbb{S})$ with $L_t(\mathbb{S})$ in the standard clustering procedure

But **can we calculate** $L_t(\mathbb{S})$?

$$L_t(\mathbb{S}) = \log \prod_{f \in \mathbb{F}_t} p(\mathbf{o}_f | \mathbb{S})$$

($\mathbb{F}_t$ is frames generated by states $\mathbb{S}_t$)

Targeted multi-accent acoustic modelling

Proposal: replace $L(\mathbb{S})$ with $L_t(\mathbb{S})$ in the standard clustering procedure

But **can we calculate** $L_t(\mathbb{S})$?

$$L_t(\mathbb{S}) = \log \prod_{f \in \mathbb{F}_t} p(\mathbf{o}_f | \mathbb{S}) \qquad (\mathbb{F}_t \text{ is frames generated by states } \mathbb{S}_t)$$

$$= \sum_{f \in \mathbb{F}_t} \log \left[\mathcal{N}(\mathbf{o}_f | \boldsymbol{\mu}(\mathbb{S}), \boldsymbol{\Sigma}(\mathbb{S})) \right] \qquad (\textbf{Gaussian} \text{ observation PDFs})$$

Targeted multi-accent acoustic modelling

Proposal: replace $L(\mathbb{S})$ with $L_t(\mathbb{S})$ in the standard clustering procedure

But **can we calculate** $L_t(\mathbb{S})$?

$$
\begin{aligned}
L_t(\mathbb{S}) &= \log \prod_{f \in \mathbb{F}_t} p(\mathbf{o}_f|\mathbb{S}) && (\mathbb{F}_t \text{ is frames generated by states } \mathbb{S}_t) \\
&= \sum_{f \in \mathbb{F}_t} \log \left[\mathcal{N}(\mathbf{o}_f|\boldsymbol{\mu}(\mathbb{S}), \boldsymbol{\Sigma}(\mathbb{S}))\right] && (\textbf{Gaussian} \text{ observation PDFs}) \\
&= -\frac{1}{2} N_t \left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|]\right\} - \frac{1}{2} n(N_x + N_t) \\
&\quad + \frac{1}{2} \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbb{S}) N_x \left[\boldsymbol{\Sigma}(\mathbb{S}_x) + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right]\right\}
\end{aligned}
$$

Targeted multi-accent acoustic modelling

Proposal: replace $L(\mathbb{S})$ with $L_t(\mathbb{S})$ in the standard clustering procedure

But **can we calculate** $L_t(\mathbb{S})$?

$$\begin{aligned}
L_t(\mathbb{S}) &= \log \prod_{f \in \mathbb{F}_t} p(\mathbf{o}_f|\mathbb{S}) && (\mathbb{F}_t \text{ is frames generated by states } \mathbb{S}_t) \\
&= \sum_{f \in \mathbb{F}_t} \log \left[\mathcal{N}(\mathbf{o}_f|\boldsymbol{\mu}(\mathbb{S}), \boldsymbol{\Sigma}(\mathbb{S}))\right] && (\textbf{Gaussian} \text{ observation PDFs}) \\
&= -\frac{1}{2} N_t \left\{\log[(2\pi)^n|\boldsymbol{\Sigma}(\mathbb{S})|]\right\} - \frac{1}{2} n (N_x + N_t) \\
&\quad + \frac{1}{2} \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbb{S}) N_x \left[\boldsymbol{\Sigma}(\mathbb{S}_x) + (\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))(\boldsymbol{\mu}(\mathbb{S}_x) - \boldsymbol{\mu}(\mathbb{S}))^{\mathrm{T}}\right]\right\}
\end{aligned}$$

Since $\boldsymbol{\mu}(\mathbb{S})$, $\boldsymbol{\mu}(\mathbb{S}_x)$, $\boldsymbol{\Sigma}(\mathbb{S})$ and $\boldsymbol{\Sigma}(\mathbb{S}_x)$ are **calculable** from only the the means and covariance matrices of the states in the corresponding clusters, the calculation of $L_t(\mathbb{S})$ for each possible cluster split is **computationally tractable**.

Targeted multi-accent acoustic modelling

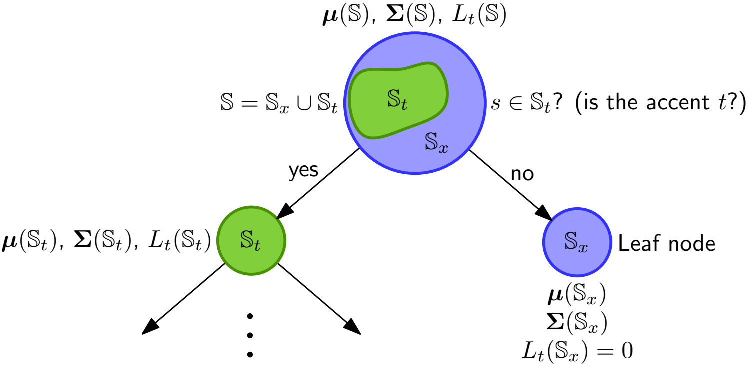
So let us take $L_t(\mathbb{S})$ as **splitting criterion** in our decision-trees

Targeted multi-accent acoustic modelling

So let us take $L_t(\mathbb{S})$ as **splitting criterion** in our decision-trees ... **problems**?

Targeted multi-accent acoustic modelling

So let us take $L_t(\mathbb{S})$ as **splitting criterion** in our decision-trees ... **problems**?



Targeted decision-tree state clustering

Phone recognition accuracy (%)

Approach	AE	BE	CE	EE	IE	Average
Accent-specific	64.80	56.77	64.59	72.97	64.27	64.68
Accent-independent	65.97	55.98	66.51	74.45	64.40	65.44
Multi-accent	66.20	56.56	66.31	73.94	64.60	65.50
Targeted multi-accent	64.60	55.17	64.11	72.65	64.44	**64.21**

Weighted targeted decision-tree state clustering

Let us **weigh** the likelihoods: $L_w(\mathbb{S}) = w_t L_t(\mathbb{S}) + w_x L_x(\mathbb{S})$

Weighted targeted decision-tree state clustering

Let us **weigh** the likelihoods: $L_w(\mathbb{S}) = w_t L_t(\mathbb{S}) + w_x L_x(\mathbb{S})$

Phone recognition accuracy (%)

Approach	AE	BE	CE	EE	IE	Average
Accent-specific	64.80	56.77	64.59	72.97	64.27	64.68
Accent-independent	65.97	55.98	66.51	74.45	64.40	65.44
Multi-accent	66.20	56.56	66.31	73.94	64.60	65.50
Targeted multi-accent	64.60	55.17	64.11	72.65	64.44	64.21
Weighted targeted	66.74	56.56	66.13	73.94	64.96	**65.65**
Weight w_t used above	0.51	0.5	0.53	0.5	0.54	

Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit **optimisation** on a **target accent**
- Showed that when likelihood is calculated **only** on **target accent**, **performance deteriorates** (possibly due to high separation of target)
- Showed that when some **weight** is also assigned to **non-target accents** (giving control over separation) very **small improvements** can be obtained
- **Criticism**: clustering early on in model training process, no guarantees
- **Future**: compare/incorporate to/in classic **adaptation** approaches

Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit **optimisation** on a **target accent**
- Showed that when likelihood is calculated **only** on **target accent**, **performance deteriorates** (possibly due to high separation of target)
- Showed that when some **weight** is also assigned to **non-target accents** (giving control over separation) very **small improvements** can be obtained
- **Criticism**: clustering early on in model training process, no guarantees
- **Future**: compare/incorporate to/in classic **adaptation** approaches

Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit **optimisation** on a **target accent**
- Showed that when likelihood is calculated **only** on **target accent**, **performance deteriorates** (possibly due to high separation of target)
- Showed that when some **weight** is also assigned to **non-target accents** (giving control over separation) very **small improvements** can be obtained
- **Criticism**: clustering early on in model training process, no guarantees
- **Future**: compare/incorporate to/in classic **adaptation** approaches

Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit **optimisation** on a **target accent**
- Showed that when likelihood is calculated **only** on **target accent**, **performance deteriorates** (possibly due to high separation of target)
- Showed that when some **weight** is also assigned to **non-target accents** (giving control over separation) very **small improvements** can be obtained
- **Criticism**: clustering early on in model training process, no guarantees
- **Future**: compare/incorporate to/in classic **adaptation** approaches

Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit **optimisation** on a **target accent**
- Showed that when likelihood is calculated **only** on **target accent**, **performance deteriorates** (possibly due to high separation of target)
- Showed that when some **weight** is also assigned to **non-target accents** (giving control over separation) very **small improvements** can be obtained
- **Criticism**: clustering early on in model training process, no guarantees
- **Future**: compare/incorporate to/in classic **adaptation** approaches