Resource development and experiments in automatic South African broadcast news transcription

SLTU 2012, Cape Town, South Africa

Herman Kamper[1], Febe de Wet[1,2], Thomas Hain[3], Thomas Niesler[1]

[1]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2]Human Language Technology Competency Area, CSIR Meraka Institute, Pretoria, South Africa
[3]Department of Computer Science, University of Sheffield, United Kingdom

UNIVERSITEIT STELLENBOSCH UNIVERSITY

The University Of Sheffield.

Introduction

Broadcast news domain:

- Provides a ready **source** of speech audio data
- Variety of speech styles and quality: careful newsreader to noisy spontaneous
- Useful as **components** for subsequent speech technologies

Introduction

Broadcast news domain:

- Provides a ready **source** of speech audio data
- Variety of speech styles and quality: careful newsreader to noisy spontaneous
- Useful as **components** for subsequent speech technologies

South African (English) broadcast news:

- Several prevalent English accents
- South African English is under-resourced variety of English

Introduction

Broadcast news domain:

- Provides a ready **source** of speech audio data
- Variety of speech styles and quality: careful newsreader to noisy spontaneous
- Useful as **components** for subsequent speech technologies

South African (English) broadcast news:

- Several prevalent English accents
- South African English is under-resourced variety of English

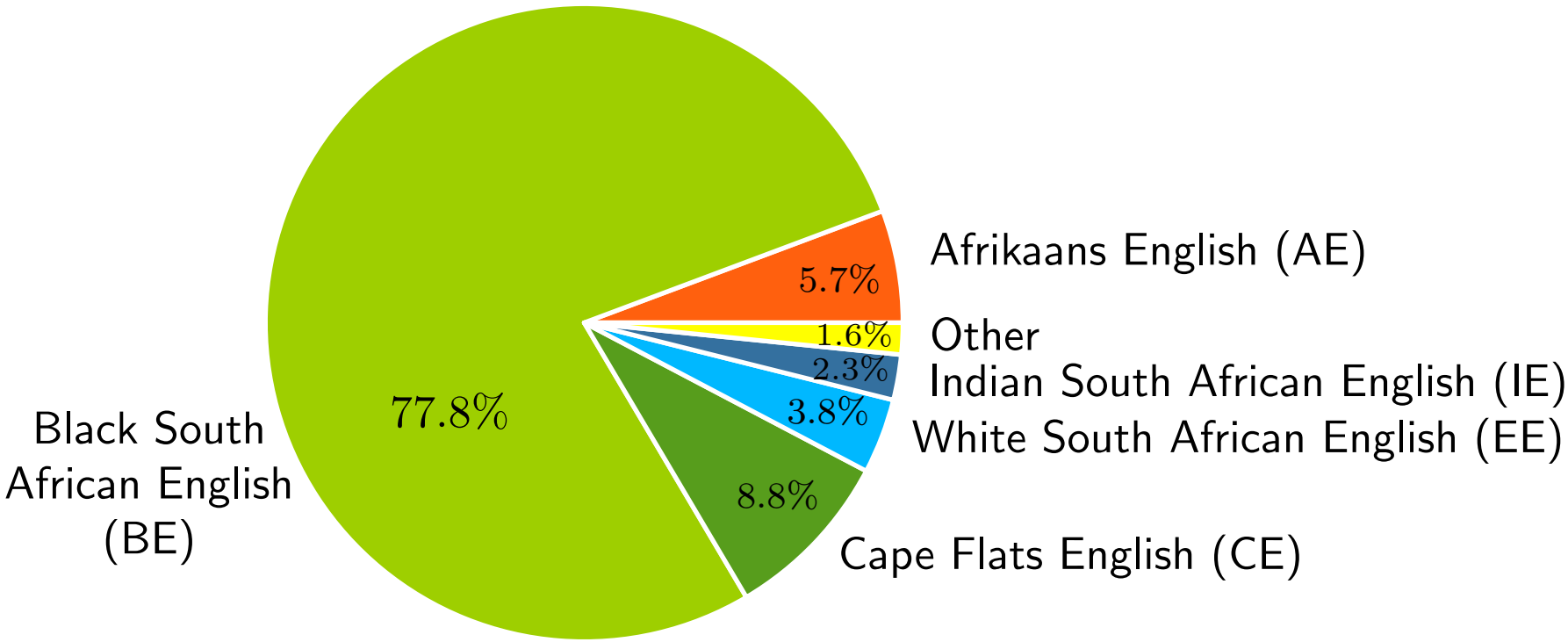
Motivation

Report on **baseline results** of a straight-forward system:

- Use resources collected at Stellenbosch University (2000 – present)
- Aim is to use baseline for comparative/interesting **further studies**

Accents of English in South Africa

Five major **accents of South African English** are identified in the literature:



South African broadcast news data

20 hours SAFM broadcasts from 1996 to 2006:

- **RD**: Newsreader speech, prepared
  27 speakers, 12.9 hours (BE, EE, IE)
- **SI**: Studio interview speech, fairly spont.
  61 speakers, 0.6 hours
- **NST**: Non-studio telephone speech, spont.
  262 speakers, 2.07 hours
- **NS**: Non-studio wideband speech, noisy
  208 speakers, 1.54 hours

Accent annotated for each sentence-level segment. Test set similar in composition to training set $\sim$2.7 hours.

System development

Speech recognition problem

$$\hat{W} = \arg\max_{W} P(W|\mathbf{X}) = \arg\max_{W} p(\mathbf{X}|W)\, P(W)$$

System development

Speech recognition problem

$$\hat{W} = \arg\max_{W} P(W|\mathbf{X}) = \arg\max_{W} p(\mathbf{X}|W)\, P(W)$$

Models required

1. Language model for $P(W)$ - 109M word corpus of newspaper text

System development

Speech recognition problem

$$\hat{W} = \arg\max_{W} P(W|\mathbf{X}) = \arg\max_{W} p(\mathbf{X}|W)\, P(W)$$

Models required

1. Language model for $P(W)$ - 109M word corpus of newspaper text
2. Pronunciation dictionary for $p(\mathbf{X}|W)$ - 60k word pronunciation dictionary

System development

Speech recognition problem

$$\hat{W} = \arg\max_{W} P(W|\mathbf{X}) = \arg\max_{W} p(\mathbf{X}|W)\, P(W)$$

Models required

1. Language model for $P(W)$ - 109M word corpus of newspaper text
2. Pronunciation dictionary for $p(\mathbf{X}|W)$ - 60k word pronunciation dictionary
3. Acoustic model for $p(\mathbf{X}|W)$ - 20h SABN corpus (previous slide)

Language modelling

- 109M word corpus from **South African newspapers**, collected 2000 – 2005: *The Financial Mail*, *Business Day*, *The Sunday Times*, *The Times*, *Sunday World*, *The Sowetan*, *The Herald*, *The Algoa Sun* and *The Daily Dispatch*
- SRILM toolkit used to train **trigram language models** on above text as well as on the transcriptions of acoustic training set (185k words)
- Also considered **interpolation** of the two language models

Language modelling

- 109M word corpus from **South African newspapers**, collected 2000 – 2005: *The Financial Mail*, *Business Day*, *The Sunday Times*, *The Times*, *Sunday World*, *The Sowetan*, *The Herald*, *The Algoa Sun* and *The Daily Dispatch*
- SRILM toolkit used to train **trigram language models** on above text as well as on the transcriptions of acoustic training set (185k words)
- Also considered **interpolation** of the two language models

Language model	*Perplexity*
Trained on 109M newspaper corpus	162.9
Trained on acoustic training set	328.9
Interpolation of the above two	**139.9**

Pronunciation dictionary

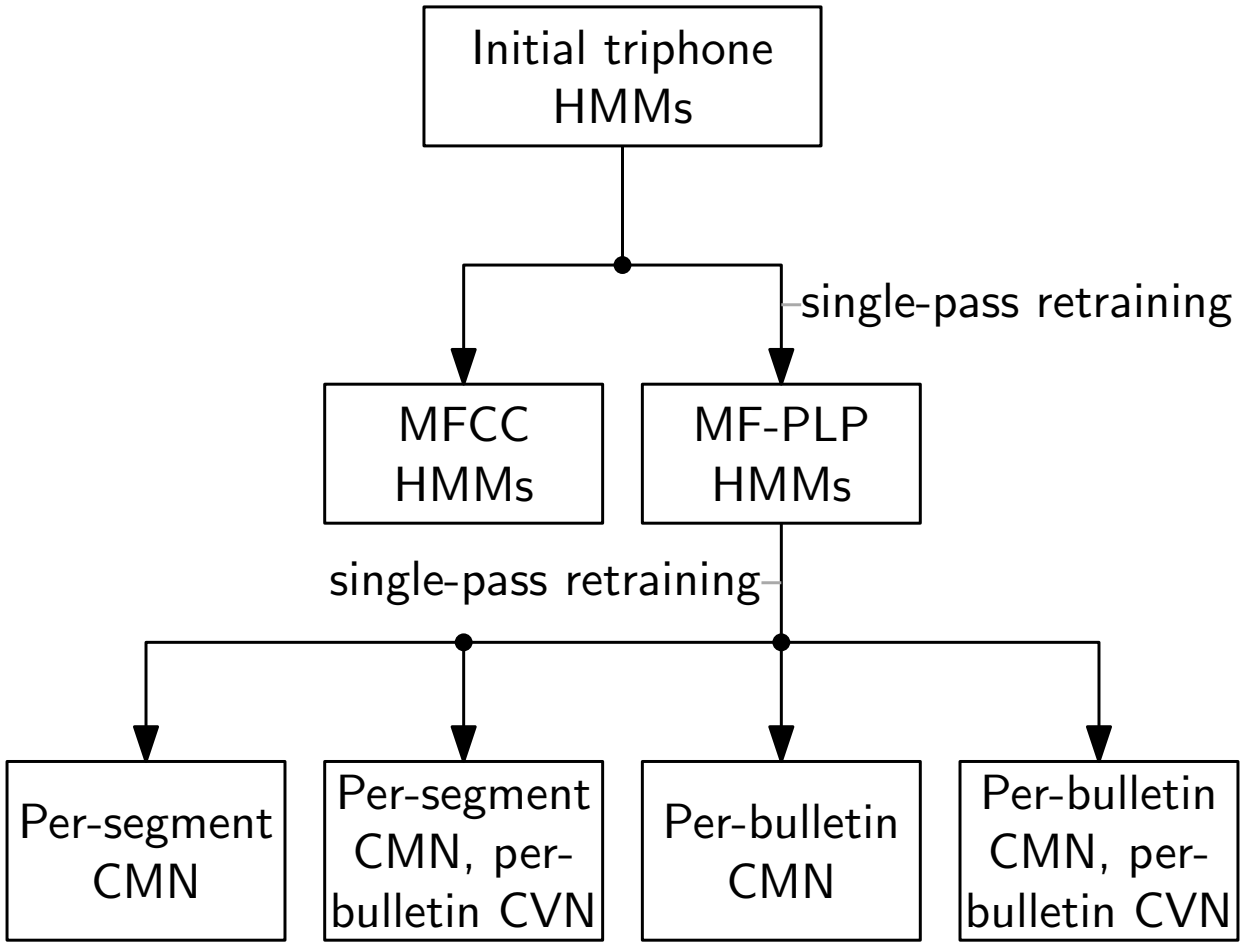
- Pronunciation dictionaries developed by a **phonetic expert**
- Reflect typical EE pronunciation
- Phone set: 45 **ARPABET** phones
- Training pronunciation dictionary: 15k words
- Recognition pronunciation dictionary: 60k words
- Average number of pronunciations per word: 1.25
- Out-of-vocabulary rate on test set: 1.02%

Acoustic modelling

Used HTK to train cross-word triphone HMMs

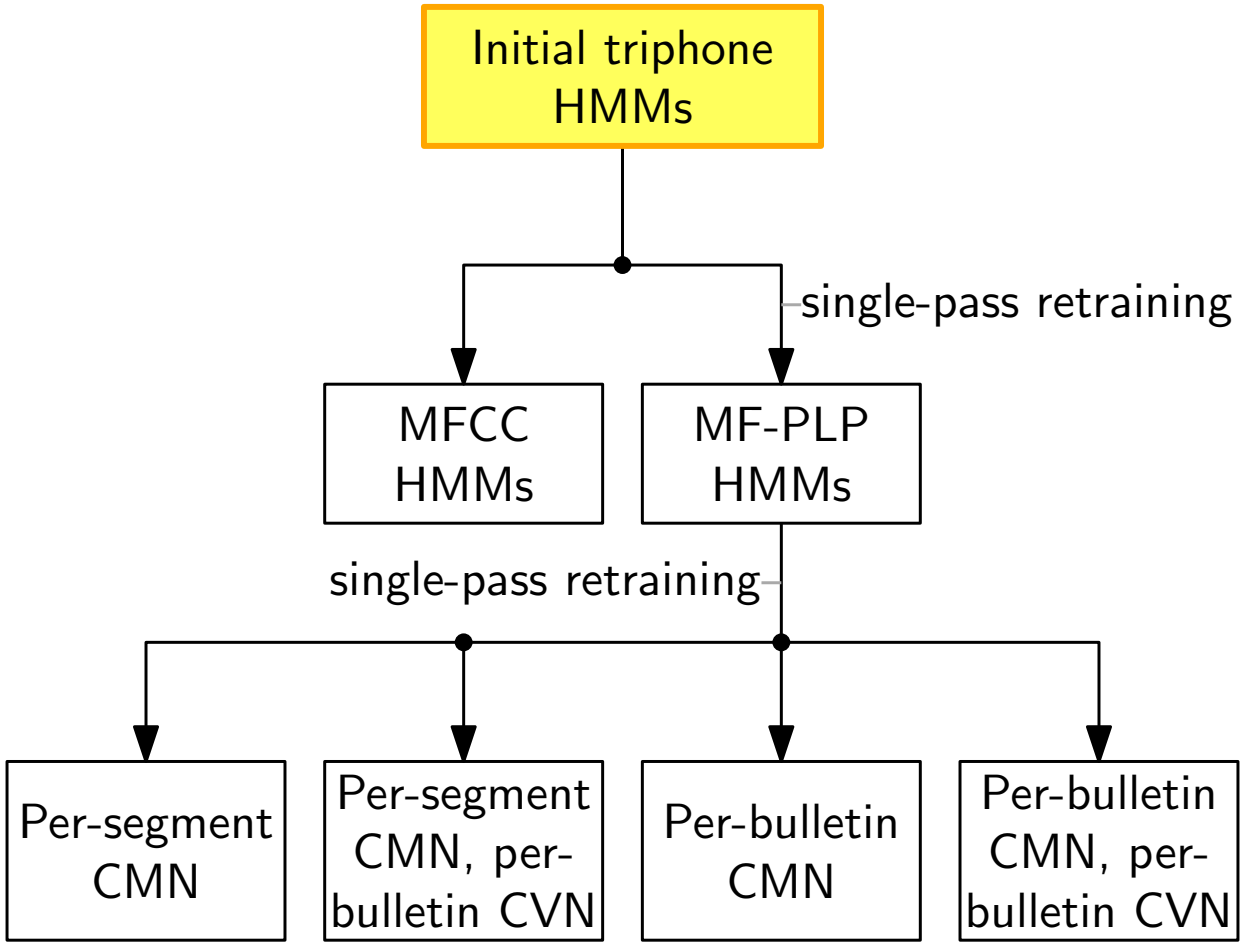
Acoustic modelling

Used HTK to train cross-word triphone HMMs



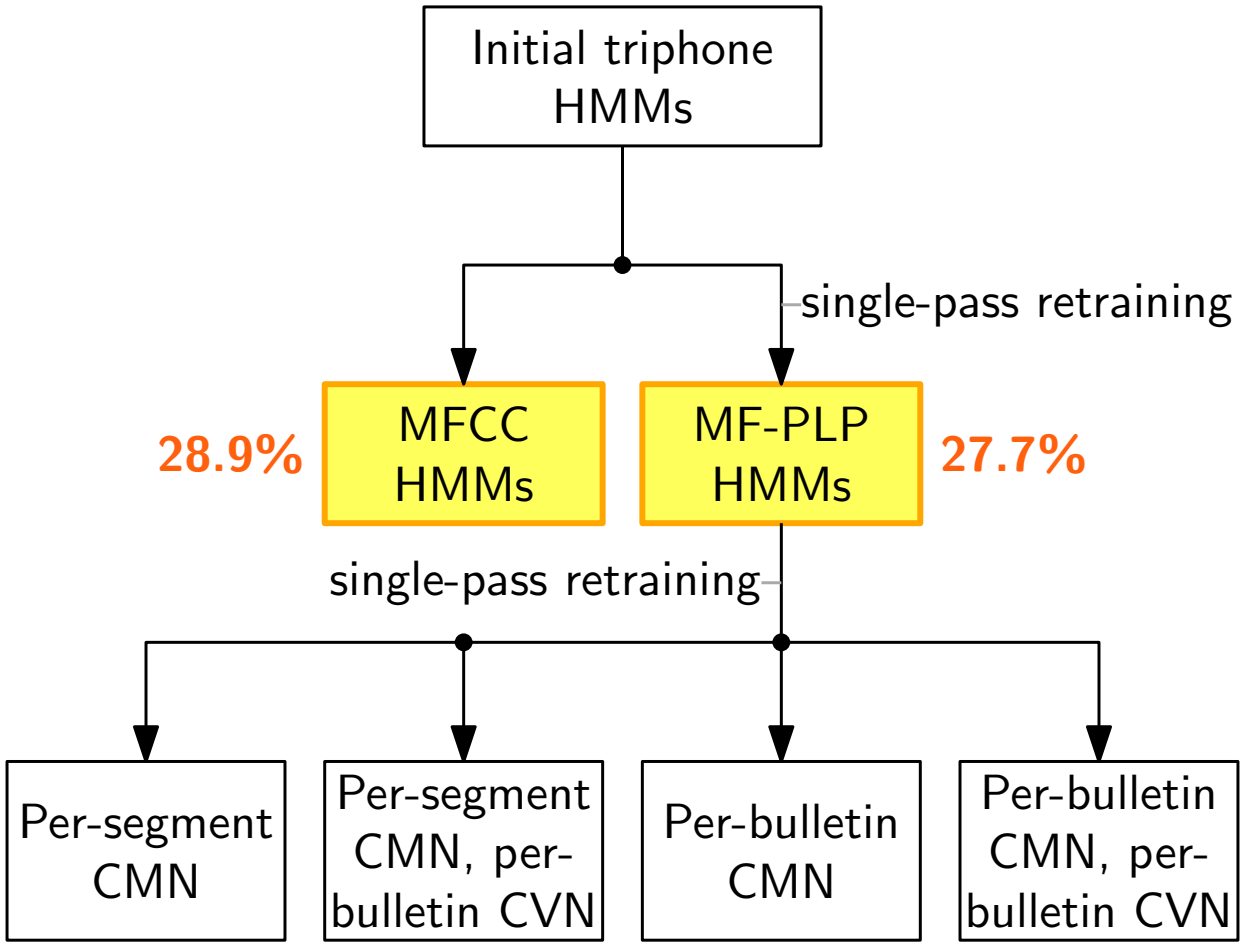
Acoustic modelling

Used HTK to train cross-word triphone HMMs



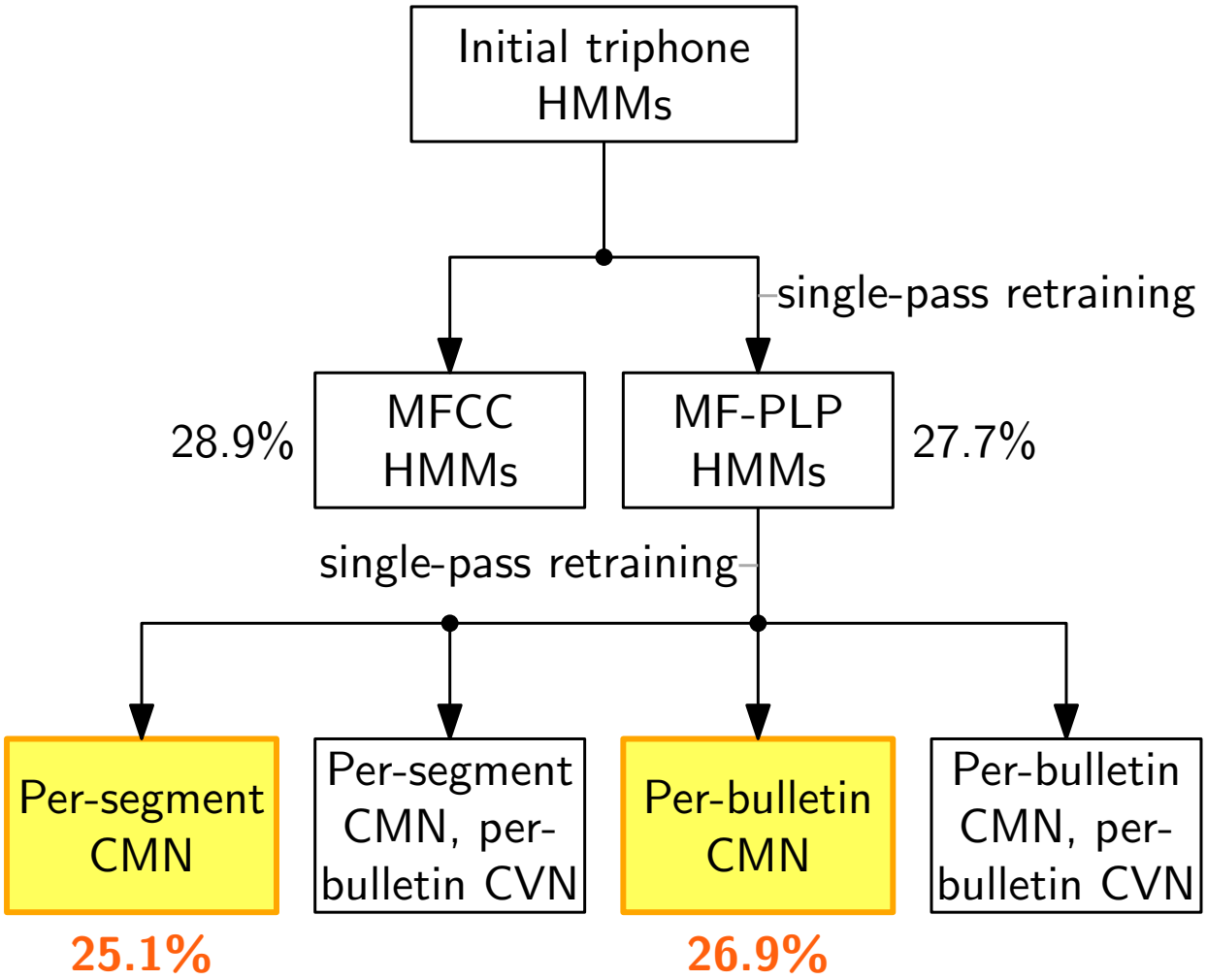
Acoustic modelling

Used HTK to train cross-word triphone HMMs



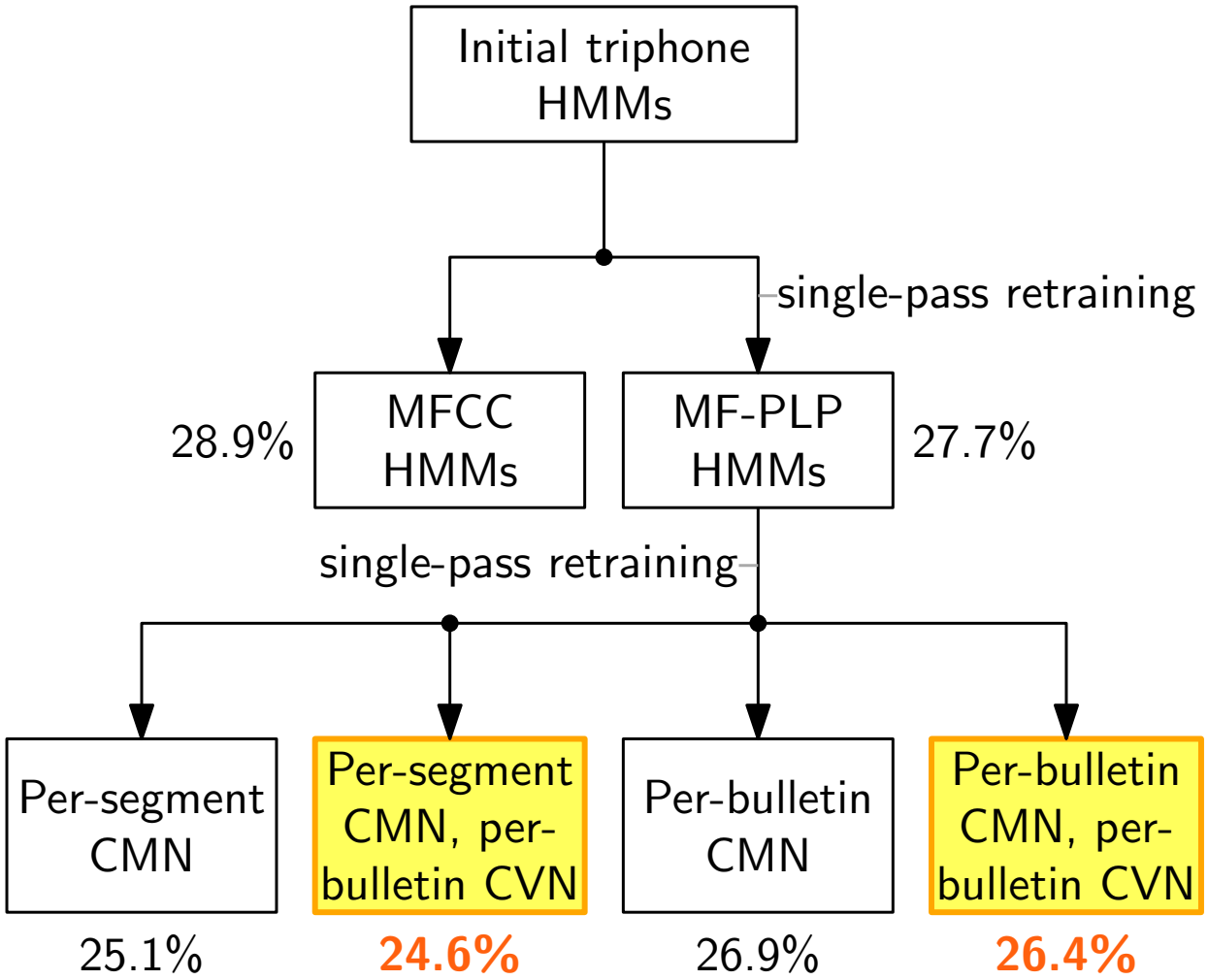
Acoustic modelling

Used HTK to train cross-word triphone HMMs



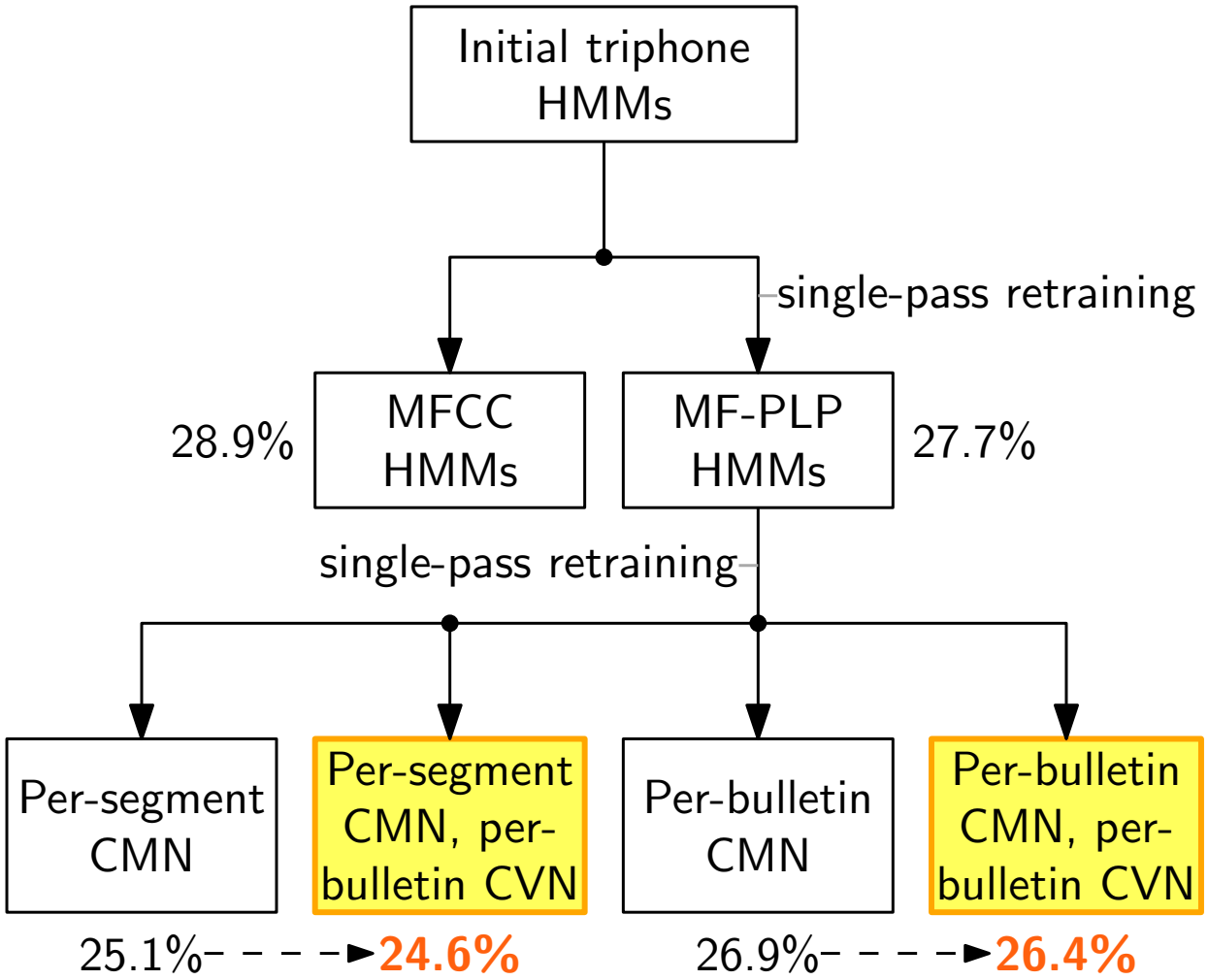
Acoustic modelling

Used HTK to train cross-word triphone HMMs



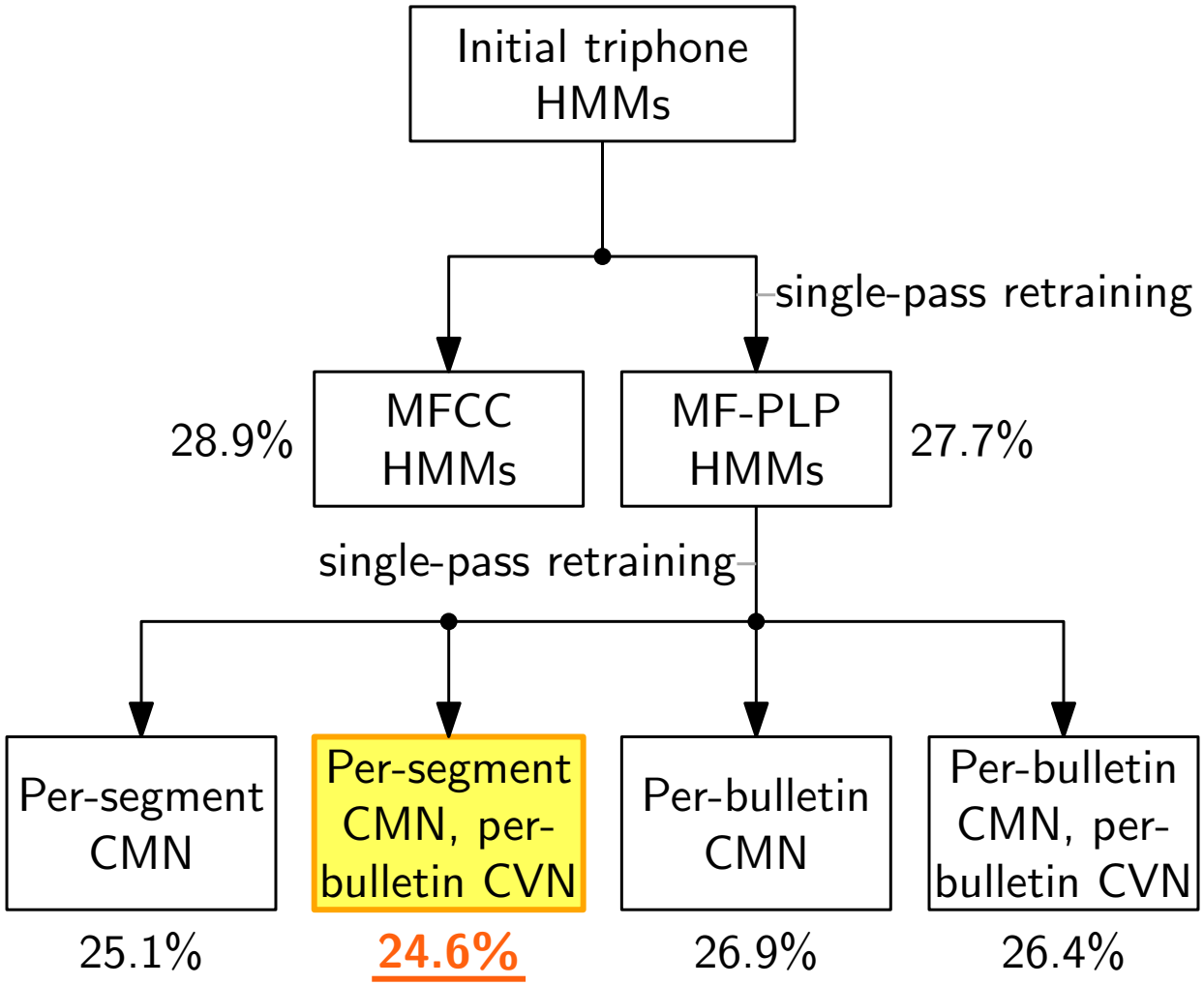
Acoustic modelling

Used HTK to train cross-word triphone HMMs



Acoustic modelling

Used HTK to train cross-word triphone HMMs



Experimental results

Final system

- Acoustic model set: 2624 states
- Features: mel-frequency perceptual linear prediction (**MF-PLP**)
- Normalisation: per-segment **CMN**, per-bulletin **CVN**

Experimental results

Final system

- Acoustic model set: 2624 states
- Features: mel-frequency perceptual linear prediction (**MF-PLP**)
- Normalisation: per-segment **CMN**, per-bulletin **CVN**

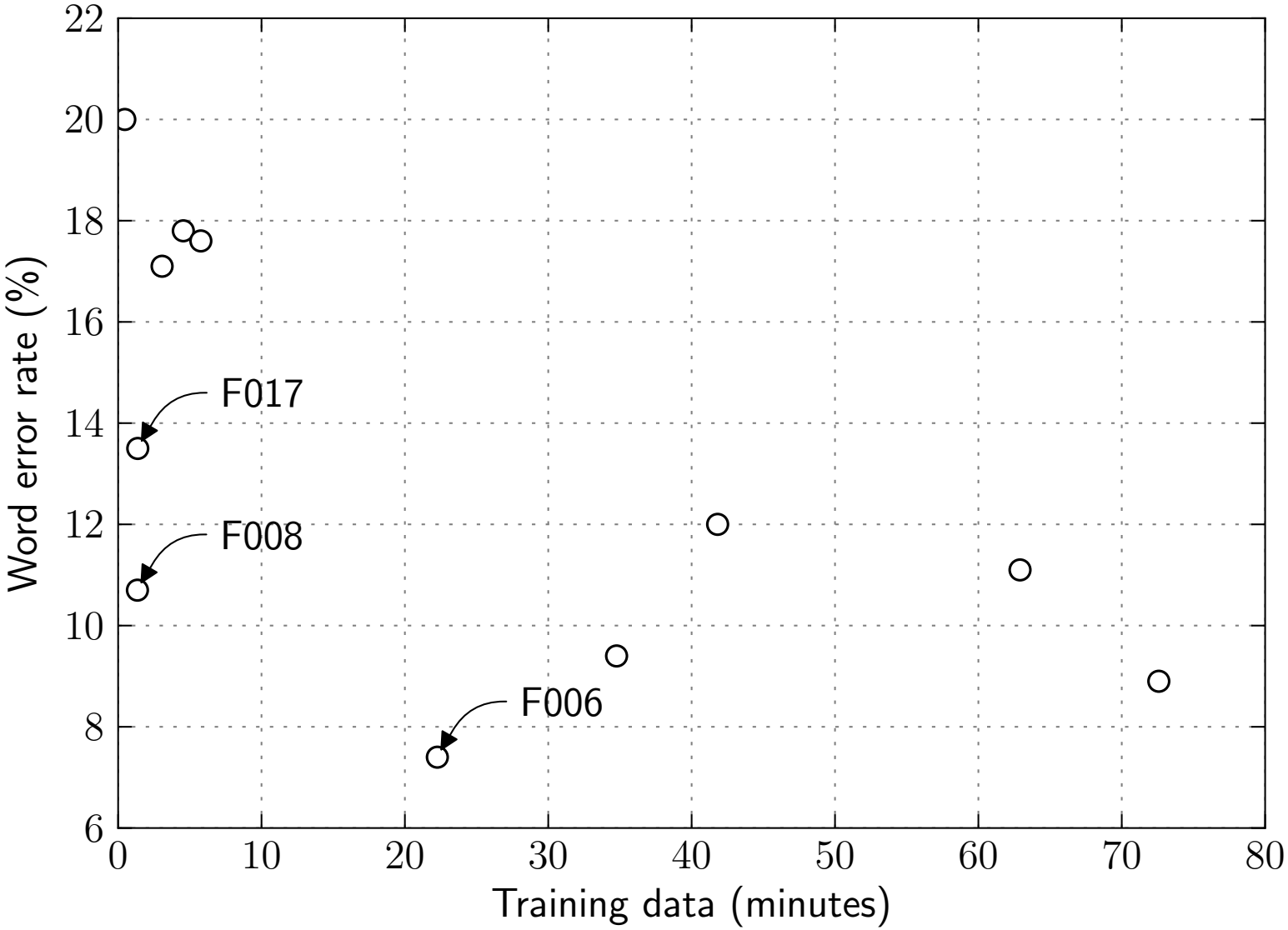
Evaluation

- Used the first-best output from **HTK HDecode** decoder
- Measured WERs separately for each **accent** and **channel condition**

System performance

Accent	*RD*	*SI*	*NST*	*NS*	*Overall*
AE	-	-	60.7	67.0	63.3
BE	13.7	19.6	64.3	56.9	29.4
CE	-	-	61.7	-	61.7
EE	14.1	-	54.1	41.6	17.2
IE	12.7	-	59.2	-	16.6
UKE	-	17.7	22.7	32.2	23.8
USE	-	39.3	-	50.5	48.0
Other	-	-	63.0	66.7	65.3
Overall	13.6	19.5	57.3	52.0	**24.6**

System performance



MP3 audio compression

MP3 bit-rate	*RD*	*SI*	*NST*	*NS*	*Overall*
128 kbps	13.6	18.9	57.0	51.9	**24.6**
64 kbps	13.4	18.8	57.8	52.3	**24.6**
32 kbps	14.3	20.8	58.7	50.7	**25.3**

Summary and conclusions

Summary:

- Described **compilation of resources** and subsequent language, pronunciation and acoustic **modelling**
- Compared MFCC and MF-PLP **parametrisation**
- **Normalisation**: compared CMN and CVN
- Considered system performance on **MP3** compressed audio

Summary and conclusions

Summary:

- Described **compilation of resources** and subsequent language, pronunciation and acoustic **modelling**
- Compared MFCC and MF-PLP **parametrisation**
- **Normalisation**: compared CMN and CVN
- Considered system performance on **MP3** compressed audio

Main findings

- Final system: MF-PLP, per-segment CMN, per-bulletin CVN
- WER of **24.6%**, poor performance on spontaneous and telephone speech
- MP3 compression: system maintains performance except at very low bit-rates

Future work

Improvements to current system:

- Presence of **several accents**: pronunciation and acoustic modelling
- Single pronunciation dictionary is currently employed

Future work

Improvements to current system:

- Presence of **several accents**: pronunciation and acoustic modelling
- Single pronunciation dictionary is currently employed

Comparative study:

- Contrast performance with similarly trained **UK** and **US** systems
- Identify how resources from well-resourced UK and US English varieties can be used in the **poorly-resourced** SA environment

Future work

Improvements to current system:

- Presence of **several accents**: pronunciation and acoustic modelling
- Single pronunciation dictionary is currently employed

Comparative study:

- Contrast performance with similarly trained **UK** and **US** systems
- Identify how resources from well-resourced UK and US English varieties can be used in the **poorly-resourced** SA environment

Comparison and substitution

SABN system	USBN system
+ SA acoustic model	+ US acoustic model
+ SA language model	+ US language model
+ SA dictionary	+ US dictionary

Future work

Improvements to current system:

- Presence of **several accents**: pronunciation and acoustic modelling
- Single pronunciation dictionary is currently employed

Comparative study:

- Contrast performance with similarly trained **UK** and **US** systems
- Identify how resources from well-resourced UK and US English varieties can be used in the **poorly-resourced** SA environment

Comparison and substitution

SABN system	USBN system
	+ US acoustic model
+ SA language model	+ US language model
+ SA dictionary	+ US dictionary

Future work

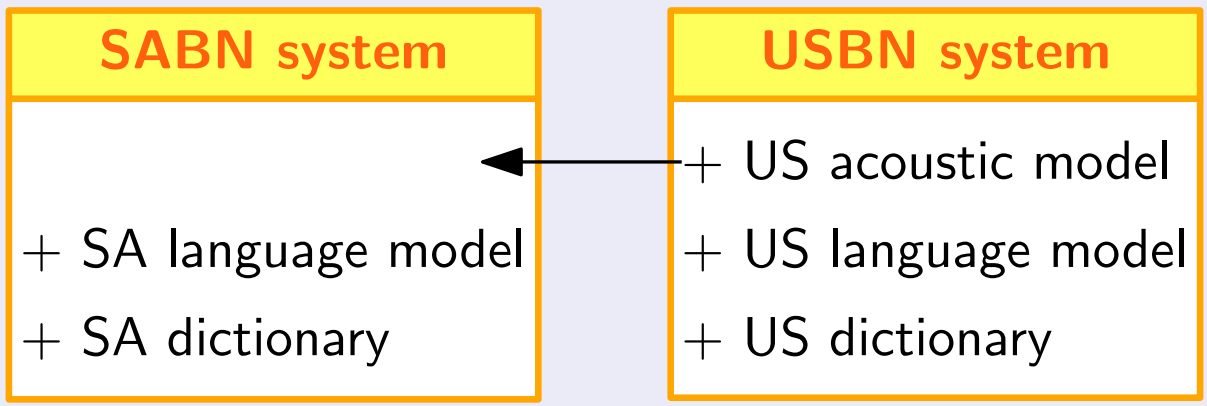
Improvements to current system:

- Presence of **several accents**: pronunciation and acoustic modelling
- Single pronunciation dictionary is currently employed

Comparative study:

- Contrast performance with similarly trained **UK** and **US** systems
- Identify how resources from well-resourced UK and US English varieties can be used in the **poorly-resourced** SA environment

Comparison and substitution



Future work

Improvements to current system:

- Presence of **several accents**: pronunciation and acoustic modelling
- Single pronunciation dictionary is currently employed

Comparative study:

- Contrast performance with similarly trained **UK** and **US** systems
- Identify how resources from well-resourced UK and US English varieties can be used in the **poorly-resourced** SA environment

Comparison and substitution

SABN system	USBN system
+ US acoustic model	
+ SA language model	+ US language model
+ SA dictionary	+ US dictionary

Future work

Improvements to current system:

- Presence of **several accents**: pronunciation and acoustic modelling
- Single pronunciation dictionary is currently employed

Comparative study:

- Contrast performance with similarly trained **UK** and **US** systems
- Identify how resources from well-resourced UK and US English varieties can be used in the **poorly-resourced** SA environment

Comparison and substitution

SABN system

+ US acoustic model
+ SA language model
+ SA dictionary

Performance penalty?