Karlsruhe Institute of Technology

The 2013 KIT Quaero Speech-to-Text System for French

Joshua Winebarger, **Bao Nguyen, Jonas Gehring, Sebastian Stüker, and Alexander Waibel**

KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT)

Quaero

Karlsruhe Institute of Technology

The Quaero Program

- Decoding Strategy
- Acoustic data & training techniques
- Feature front-end developments
  - Deep neural networks
  - Tonal features
- Pronunciation Modeling
  - Pseudographemes
- Language Modeling
- Experiments
  - Elision normalisation
  - Acoustic Data Filtering
- Results
- Conclusion

The Quaero Program

Overview

- French research program with German participation
- Centered on multilingual and multimedia prcessing
- Evaluation-driven, corpus-based
- Coopetition dynamic
  - Partners work separately on the same subjects
  - Regular exchange of findings and results

Areas of research:

- Development of access
- Automatic extraction
- Analysis
- Classification
- Application of information

Focus on:

- Innovation
- Applications

The Quaero Speech-to-Text Task

- Automatic speech recognition (ASR) is a key Quaero technology

Useful for:

- multimedia search
- content enrichment
- structuring multimedia and multilingual documents

Domain:

- Broadcast news and conversation
- Challenging due to:
  - Disfluencies
  - Non-speech events

- Fall 2013 is the fifth and final Quaero ASR evaluation

Karlsruhe Institute of Technology

System Overview

Decoding Strategy

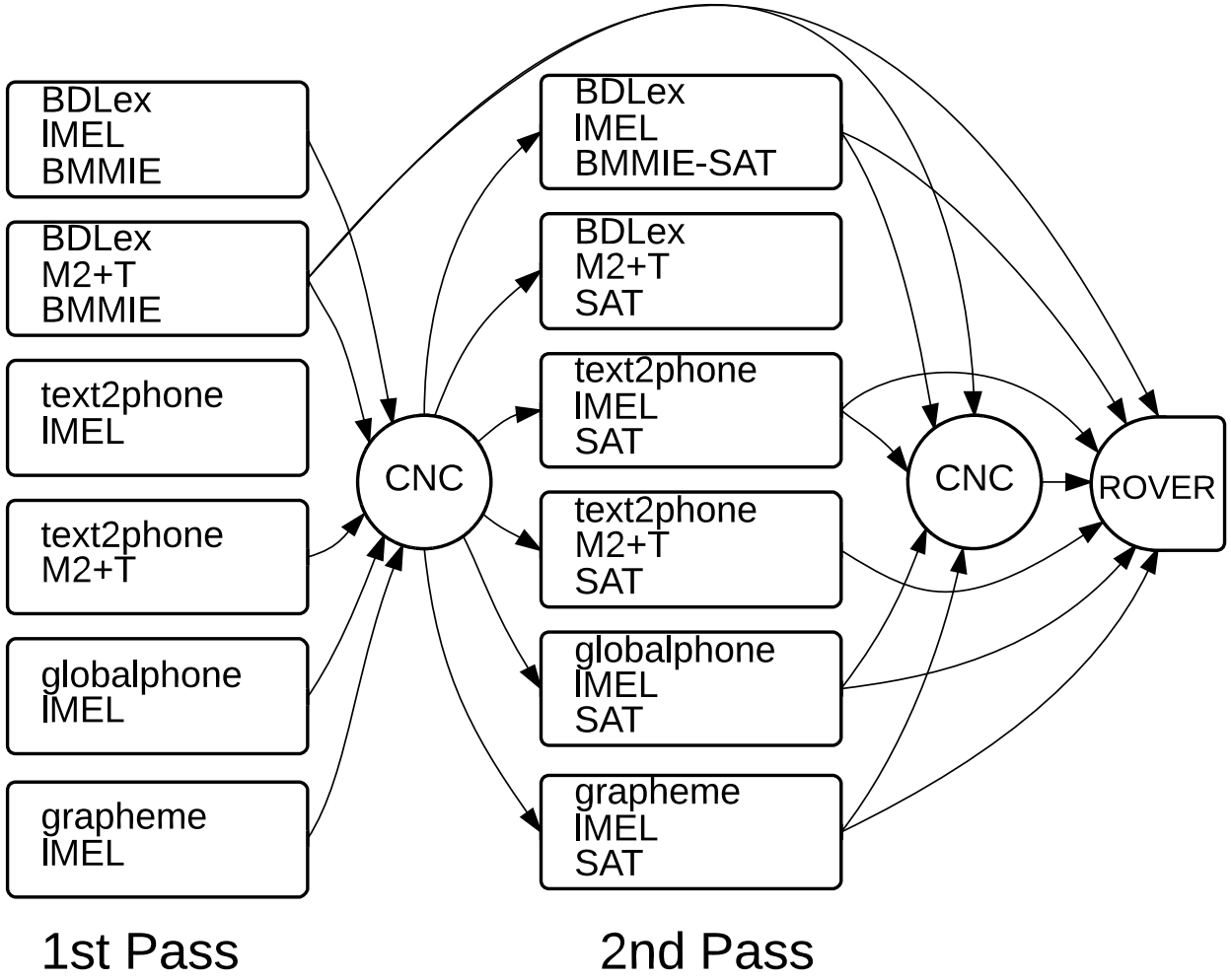


Table of Contents

Acoustic Modeling

Karlsruhe Institute of Technology

The Quaero Program

Decoding Strategy

Acoustic data & training techniques

Feature front-end developments

- Deep neural networks
- Tonal features

Pronunciation Modeling

- Pseudographemes

Language Modeling

Experiments

- Elision normalisation
- Acoustic Data Filtering

Results

Conclusion

Acoustic Modeling Overview

- **Strategy:** Multiple subsystems w. diverse pronunciation dictionaries (among them pseudographemes) and feature front-ends $\rightarrow$ system combination

  Innovations in front-ends
  - tonal features
  - deep bottleneck neural networks

  Front-end architecture
  - log mel filter bank coefficients (**IMEL**)
  - stacked combination of MFCC, MVDR and tonal features (**M2+T**)
- Acoustic modeling with HMMs
  - Generalized quinphones
  - Three states per phone without skip states
  - Train using classification and regression decision tree
- 8K acoustic models for phoneme systems
- 12K models for pseudographeme systems

AM Training & Data

Training

- Incremental splitting of Gaussians (Merge and split)
- Estimation of one global semi-tied covariance matrix after LDA
- Two iterations of viterbi training
- VTLN
- Feature-space constrained MLLR SAT models for second-pass decoding
- Discriminative training for some systems

- All subsystems: 267 hours of speech
  - Quaero 2009-2011
  - Ester
- Acoustic data filtered by decoding

Source	No filt.	Filt.
Quaero	194.1 hrs	187.7 hrs
Ester	107.8 hrs	80.3 hrs

AM Training & Data

Training

- Incremental splitting of Gaussians (Merge and split)
- Estimation of one global semi-tied covariance matrix after LDA
- Two iterations of viterbi training
- VTLN
- Feature-space constrained MLLR SAT models for second-pass decoding
- Discriminative training for some systems

- All subsystems: 267 hours of speech
  - Quaero 2009-2011
  - Ester
- Acoustic data filtered by decoding

Source	No filt.	Filt.
Quaero	194.1 hrs	187.7 hrs
Ester	107.8 hrs	80.3 hrs

AM Training & Data

Training

- Incremental splitting of Gaussians (Merge and split)
- Estimation of one global semi-tied covariance matrix after LDA
- Two iterations of viterbi training
- VTLN
- Feature-space constrained MLLR SAT models for second-pass decoding
- Discriminative training for some systems

- All subsystems: 267 hours of speech
  - Quaero 2009-2011
  - Ester
- Acoustic data filtered by decoding

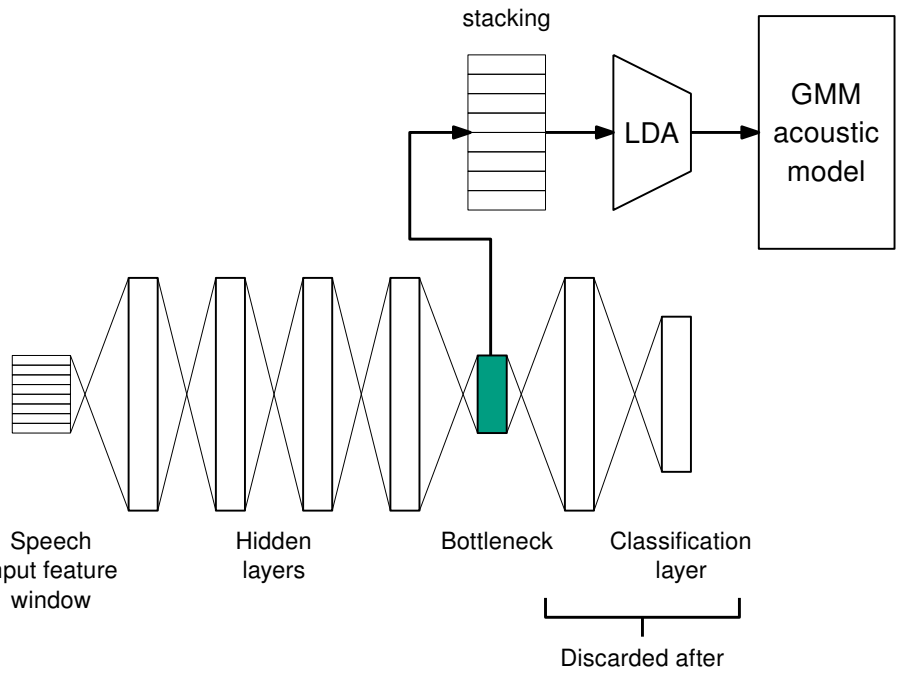
Source	**No filt.**	**Filt.**
Quaero	194.1 hrs	187.7 hrs
Ester	107.8 hrs	80.3 hrs

Karlsruhe Institute of Technology

Deep Neural Networks as Features

- Bottleneck features (BNFs) from multilayer perceptrons now common in ASR
  - Discriminative power, robustness
- We use architecture based Denoising Autoencoders (Gehring, et. al)
  - Many hidden layers
  - Significant reduction in WER ($\approx$ 20% relative)

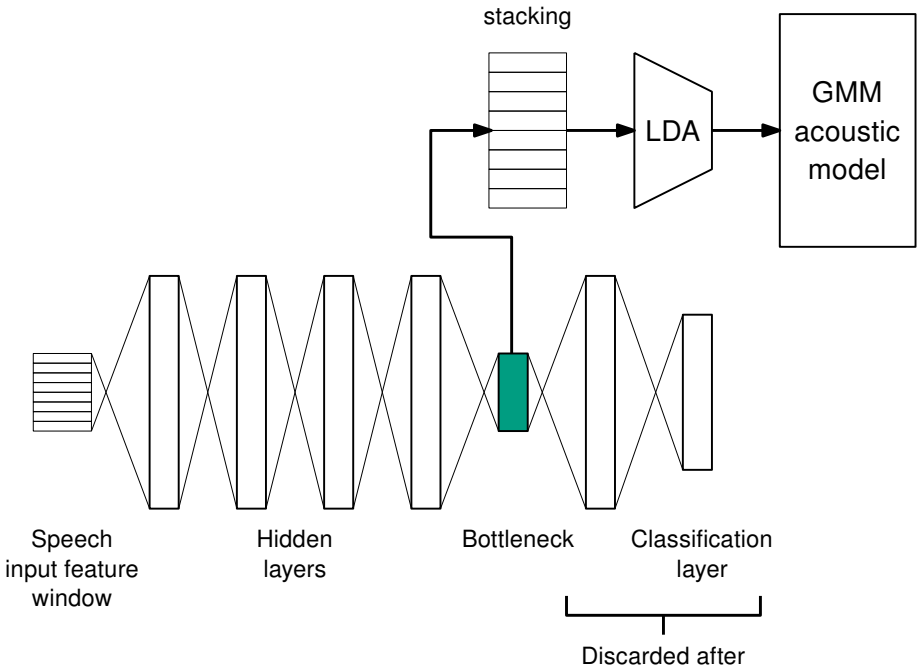
DBNF Architecture



Architecture

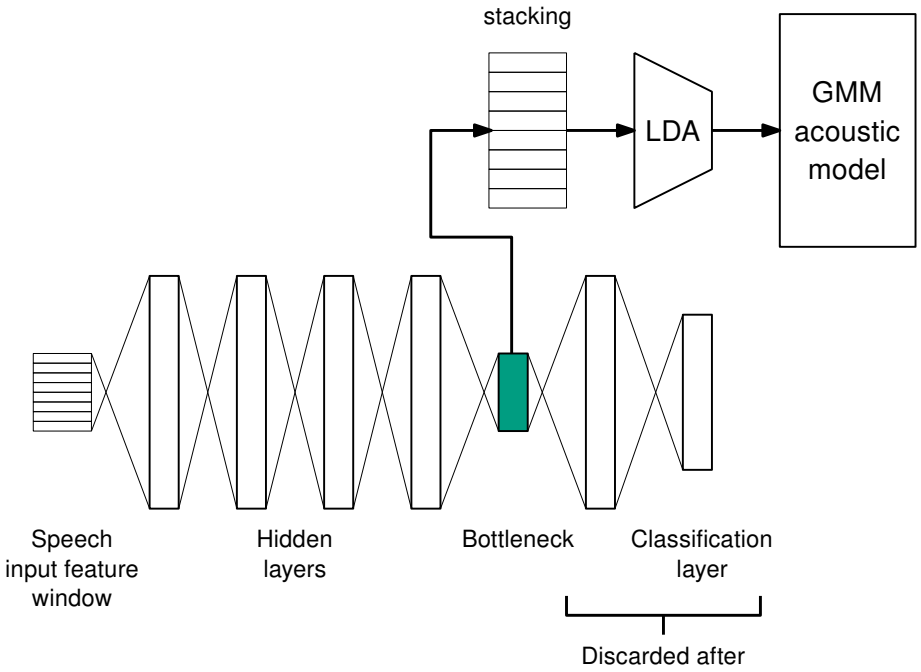
- Five fully-connected hidden layers
- 1200 units per layer
- Bottleneck layer of 42 units

DBNF Training



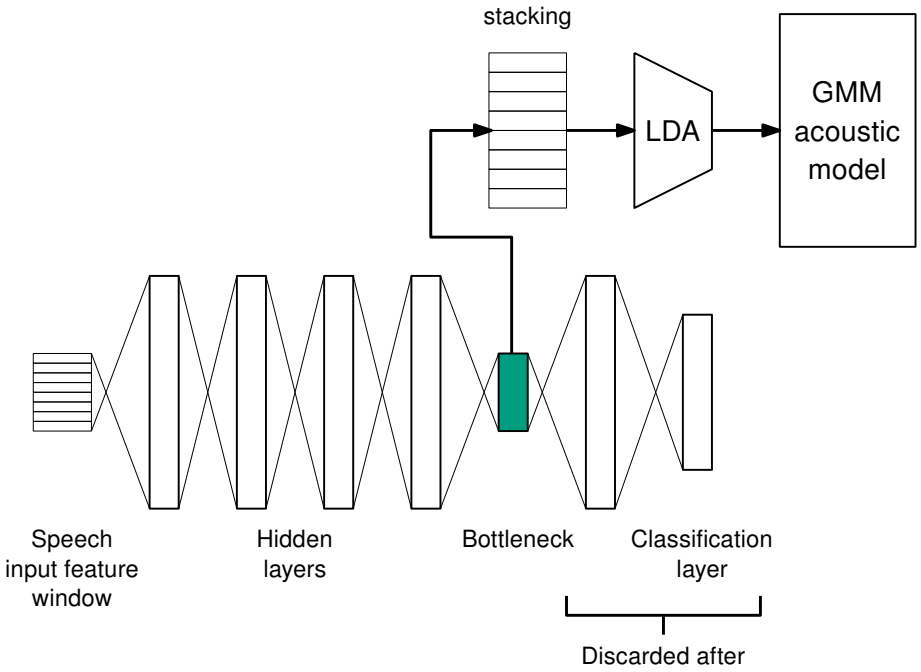
1. Hidden layers pre-trained
   - Layerwise
   - Unsupervised
   - Stack of denoising autoencoders
2. Bottleneck + next hidden layer + classification layer init.'d w. random weights
3. These three layers are connected to the autoencoders

DBNF Training



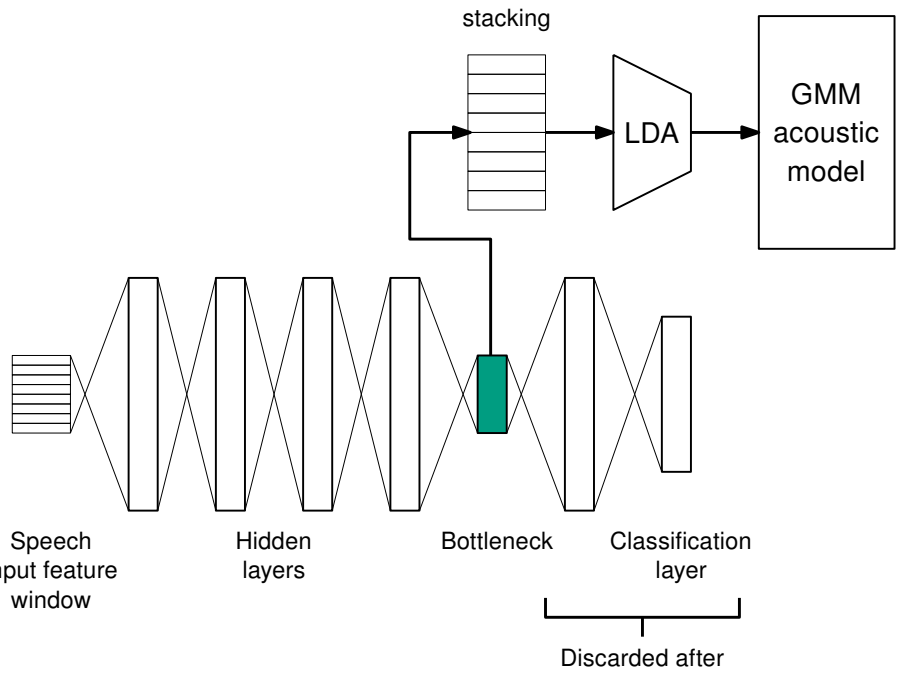
1. Hidden layers pre-trained
   - Layerwise
   - Unsupervised
   - Stack of denoising autoencoders
2. Bottleneck + next hidden layer + classification layer init.'d w. random weights
3. These three layers are connected to the autoencoders

DBNF Training



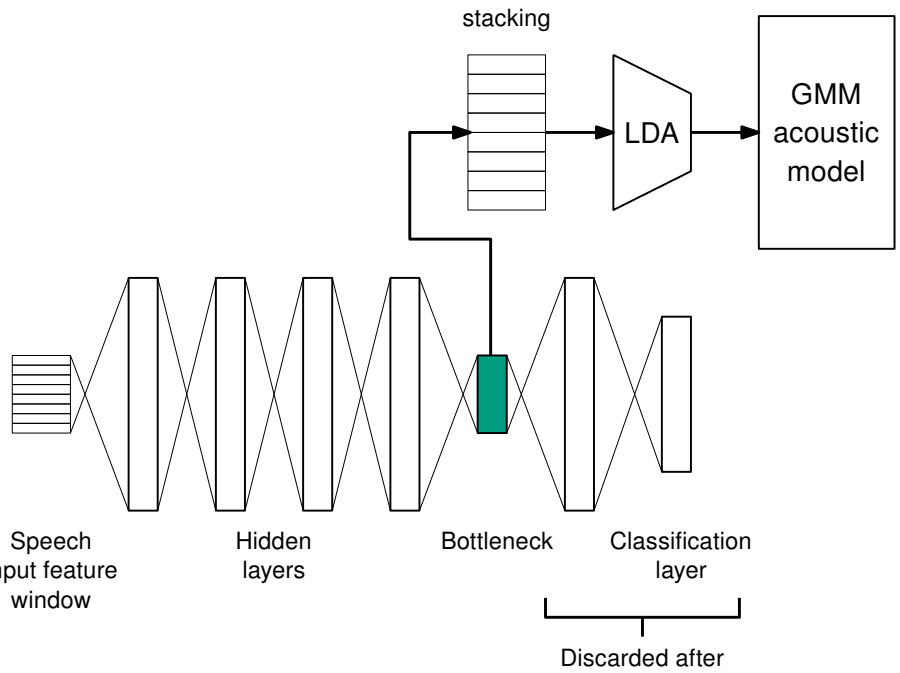
1. Hidden layers pre-trained
   - Layerwise
   - Unsupervised
   - Stack of denoising autoencoders
2. Bottleneck + next hidden layer + classification layer init.'d w. random weights
3. These three layers are connected to the autoencoders

DBNF Training



4. Entire network trained with supervision
   - Estimation of context-dependent polyphone states
   - ‘newbob’ learning rate schedule : about 14-18 epochs

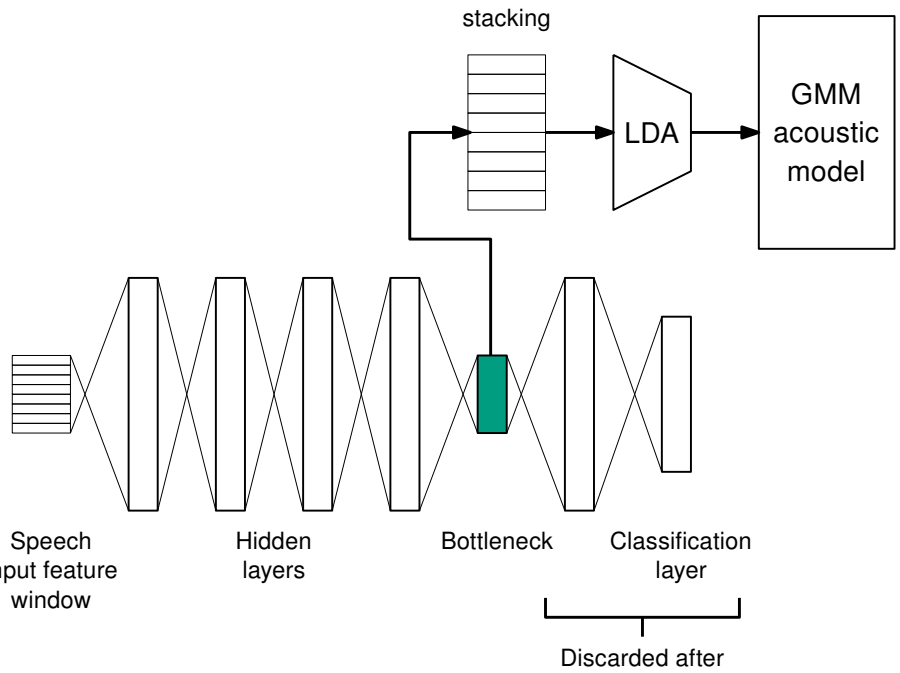
DBNF Utilisation



- Compute DBNF features
- Feature stacking
- LDA

- Relative reductions in WER (vs. MFCC):
  - **IMEL DBNF** : 22%
  - **M2+T DBNF** : 24%

DBNF Utilisation



- Compute DBNF features
- Feature stacking
- LDA

- Relative reductions in WER (vs. MFCC):
  - **IMEL DBNF** : 22%
  - **M2+T DBNF** : 24%

Tonal Features

- Tonal features traditionally thought not helpful for non-tonal languages
- Recent experiments at KIT show this to be untrue
- French M2+T-based DBNFs use two tonal features:
  - Pitch
  - Fundamental Frequency Variation (FFV)
- Tonal features bring relative improvements in WER:
  - 5% for Vietnamese (Tonal)
  - 3% for English (Non-tonal)
  - 3% for French (Non-tonal)

Tonal Features

- Tonal features traditionally thought not helpful for non-tonal languages
- Recent experiments at KIT show this to be untrue
- French M2+T-based DBNFs use two tonal features:
  - Pitch
  - Fundamental Frequency Variation (FFV)
- Tonal features bring relative improvements in WER:
  - 5% for Vietnamese (Tonal)
  - 3% for English (Non-tonal)
  - 3% for French (Non-tonal)

Tonal Features

- Tonal features traditionally thought not helpful for non-tonal languages
- Recent experiments at KIT show this to be untrue
- French M2+T-based DBNFs use two tonal features:
  - Pitch
  - Fundamental Frequency Variation (FFV)
- Tonal features bring relative improvements in WER:
  - 5% for Vietnamese (Tonal)
  - 3% for English (Non-tonal)
  - 3% for French (Non-tonal)

Tonal features

Pitch

- Cepstrogram & Autocorrelation function
- Tracks across local maxima in Dynamic-Programming way to give highest cumulative path

Fundamental Frequency Variation (FFV)

- No explicit segmentaiton into speech and silence necessary
- Computation of “vanishing point product” across range of τ
- Filterbank $\rightarrow$ 7 coefficients

Tonal features

Pitch

- Cepstrogram & Autocorrelation function
- Tracks across local maxima in Dynamic-Programming way to give highest cumulative path

Fundamental Frequency Variation (FFV)

- No explicit segmentaiton into speech and silence necessary
- Computation of “vanishing point product” across range of τ
- Filterbank $\rightarrow$ 7 coefficients

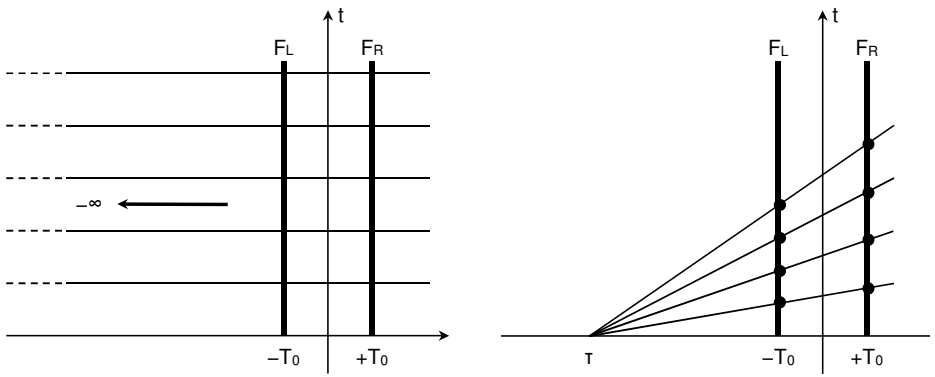


Figure: *Visualization of the vanishing point product employed in FFV.*

Karlsruhe Institute of Technology

Pronunciation modeling Overview

- Subsystems based on four pronunciation dictionaries:
  - BDLex lexicon
  - Globalphone dictionary
  - Rule-based pronunciation generation (text2phone)
  - Pseudographeme approach
- Missing pronunciations : G2P

Dictionary	WER (Clns.)
BDLex	25.4
text2phone	25.6
globalphone	26.6
pseudographeme	27.0

- AMs for BDLex & Globalphone : bootstrap from German models
- AMs for text2phone : bootstrap from BDLex models
- AMs for pseudographemes. : flatstart

Grapheme Background

- Traditional grapheme-based pronunciation modeling:
  - symbols of written word used as sub-units of pronunciation
- Using graphemes instead of phonemes is:
  - Convenient
  - Effective in reducing WER in combination w. phoneme systems
- French orthography is relatively regular
  - However no one-to-one mapping between graphemes and phones for all graphemes
  - Clusters of graphemes can produce same sounds (esp. vowels)
  - Some graphemes can be silent
- This problem can be handled by creating simple rules:
  - Treat certain groups of graphemes as the same grapheme

Grapheme Background

- Traditional grapheme-based pronunciation modeling:
  - symbols of written word used as sub-units of pronunciation
- Using graphemes instead of phonemes is:
  - Convenient
  - Effective in reducing WER in combination w. phoneme systems

- French orthography is relatively regular
  - However no one-to-one mapping between graphemes and phones for all graphemes
  - Clusters of graphemes can produce same sounds (esp. vowels)
  - Some graphemes can be silent
- This problem can be handled by creating simple rules:
  - Treat certain groups of graphemes as the same grapheme

Rule-based Approach

{ é }, { a i }, { é e }, { u é }	→	é
{ e n t }	→	ent
{ e }, { è }, { ê }	→	e
{ a u }, { e a u }	→	au
{ o n }	→	on
{ o i }	→	oin
{ g n }	→	gn

Table: *Some selected rules for merging graphemes*

- French orthography is relatively regular
  - However no one-to-one mapping between graphemes and phones for all graphemes
  - Clusters of graphmes can produce same sounds (esp. vowels)
- This problem can be handled by creating simple rules:
  - Treat certain groups of graphemes as the same grapheme

Rule-based Approach

{ é }, { a i }, { é e }, { u é }	→	é
{ e n t }	→	ent
{ e }, { è }, { ê }	→	e
{ a u }, { e a u }	→	au
{ o n }	→	on
{ o i }	→	oin
{ g n }	→	gn

Table: *Some selected rules for merging graphemes*

délaissées (*adj.* abandoned)	→	d é l é s é s
faisceaux (*n.* bundles)	→	f é s c {au} x
foncièrement (*adv.* fundamentally)	→	f {on} c i e r e m {ent}
joignent (*v.* join)	→	j {oi} {gn} {ent}
pointée (*adj.* pointed)	→	p {oin} t é

Table: *Selected entries from the grapheme dictionary with accompanying English translations*

Language Modeling Development

- 4-gram case-sensitive LM
- Modified Kneser-Ney smoothing
- Individual LMs built for each text source
- LMs interpolated using weights estimated on tuning set

Sources

Quaero transcripts, other Quaero sources
Gigaword
Europarl transcripts
Other newspapers
Assorted small oral corpora
CFPP 2000

Experiments

Background on Elision

- Elision: Final unstressed vowel immediately before another vowel-beginning word *sometimes* dropped and both words joined

Background on Elision

- Elision: Final unstressed vowel immediately before another vowel-beginning word *sometimes* dropped and both words joined

- Examples of elision:

“la	apparance”	→	l’apparance	(“the appearance”)
DEF-ART	NOUN			
“de	être”	→	d’être	(“to be”)
PREP	VERB			
“je	aimerais”	→	j’aimerais	(“I would like”)
SUBJ-PRONOUN	VERB			

Background on Elision

- Elision: Final unstressed vowel immediately before another vowel-beginning word *sometimes* dropped and both words joined

- Examples of elision:

“la	apparance”	→	l’apparance	(“the appearance”)
DEF-ART	NOUN			
“de	être”	→	d’être	(“to be”)
PREP	VERB			
“je	aimerais”	→	j’aimerais	(“I would like”)
SUBJ-PRONOUN	VERB			

- However note:

“ça	arrive”	(“that happens”)
DEM-PRONOUN	VERB	
“déjà	estimé	(“already estimated”)
ADV	VERB	

Problem & Analysis

Elision poses a challenge depending on how we consider elided words:

- *"join"* : Elisions are one token. → Large OOV, limited power of vocabulary
- *"sep"* : Elisions are two tokens. → Reduces OOV but decreases language and polyphone context
- *"sep. topN"* : Treat top *N* most common elisions as one token

Quaero transcripts

	# Tokens	Total	Elided
Total words	23M	100%	-
Elided words	165K	7.2%	100%
• top50 words	100K	4.4%	61%
• non-top50 words	65K	2.8%	39%

Problem & Analysis

Karlsruhe Institute of Technology

Elision poses a challenge depending on how we consider elided words:

- *"join"* : Elisions are one token. $\rightarrow$ Large OOV, limited power of vocabulary
- *"sep"* : Elisions are two tokens. $\rightarrow$ Reduces OOV but decreases language and polyphone context
- *"sep. topN"* : Treat top *N* most common elisions as one token

Quaero transcripts

	# Tokens	Total	Elided
Total words	23M	100%	-
Elided words	165K	7.2%	100%
• top50 words	100K	4.4%	61%
• non-top50 words	65K	2.8%	39%

Results

- Treat AM & LM training data to reflect these approaches
- Compute resulting OOV and WER

Elision treatment	OOV	WER(CI)
join	5.0%	29.6%
sep.	0.61%	26.7%
sep. w/ top50	0.68%	26.2%

Table: *OOV rate for 250K-word vocabularies on eval2011 data.*

Filtering Background & Approach

- Initial system prototyping done around small (140 hrs) Quaero data
- Additional Quaero and Ester data did not initially improve performance
- It was decided to filter the additional acoustic training data

Training steps

1. Divide training data into six parts
   1. Quaero-core (140hrs) (2010, 2011) (unfiltered)
   2. Quaero-fast (2009, 2010)
   3. Quaero-careful (2009, 2010)
   4. Ester1
   5. Ester2
   6. Ester2-dev
2. Decode on each part
3. Compute WER for each utterance
4. Reject utterances w. WER $> X\%$
5. Retrain

Filtering Background & Approach

- Initial system prototyping done around small (140 hrs) Quaero data
- Additional Quaero and Ester data did not initially improve performance
- It was decided to filter the additional acoustic training data

Training steps

1. Divide training data into six parts
   1. Quaero-core (140hrs) (2010, 2011) (unfiltered)
   2. Quaero-fast (2009, 2010)
   3. Quaero-careful (2009, 2010)
   4. Ester1
   5. Ester2
   6. Ester2-dev
2. Decode on each part
3. Compute WER for each utterance
4. Reject utterances w. WER $> X\%$
5. Retrain

Choosing a Threshold

- How do we choose X%?

Quaero data

- Reject those top 10% of utterances with highest WER
  - $\rightarrow$ reject those with WER $>$ 75%
- As alternative, reject those top 25%
  - $\rightarrow$ reject those with WER $>$ 50%

Ester data

- Ester data is different from Quaero data
  - BN vs. BC
  - Telephone and/or strongly accented speech
  - $\rightarrow$ Reject more utterances (WER $>$ 37%)

Results

Effect on WER

Training material	WER (CI)
baseline (Quaero core)	27.0
baseline + unfiltered add. Quaero & Ester	29.0
baseline + filtered add. Quaero (75%)	26.9
baseline + filtered add. Quaero (50%)	26.5
baseline + filt'd add. Q. (50%) + Ester 1, 2, & dev (37%)	26.3

Effect on DB Size

Source	Utts. (K)			Hours		
	Unfilt.'d	Kept	%	Unfilt.'d	Kept	%
Quaero core	32.5	32.5	100	140	140	100
add. Quaero + Ester	31.7	18.1	56.9	162	128	78.9
Total	64.2	50.6	78.72	302	267	88.7

Results

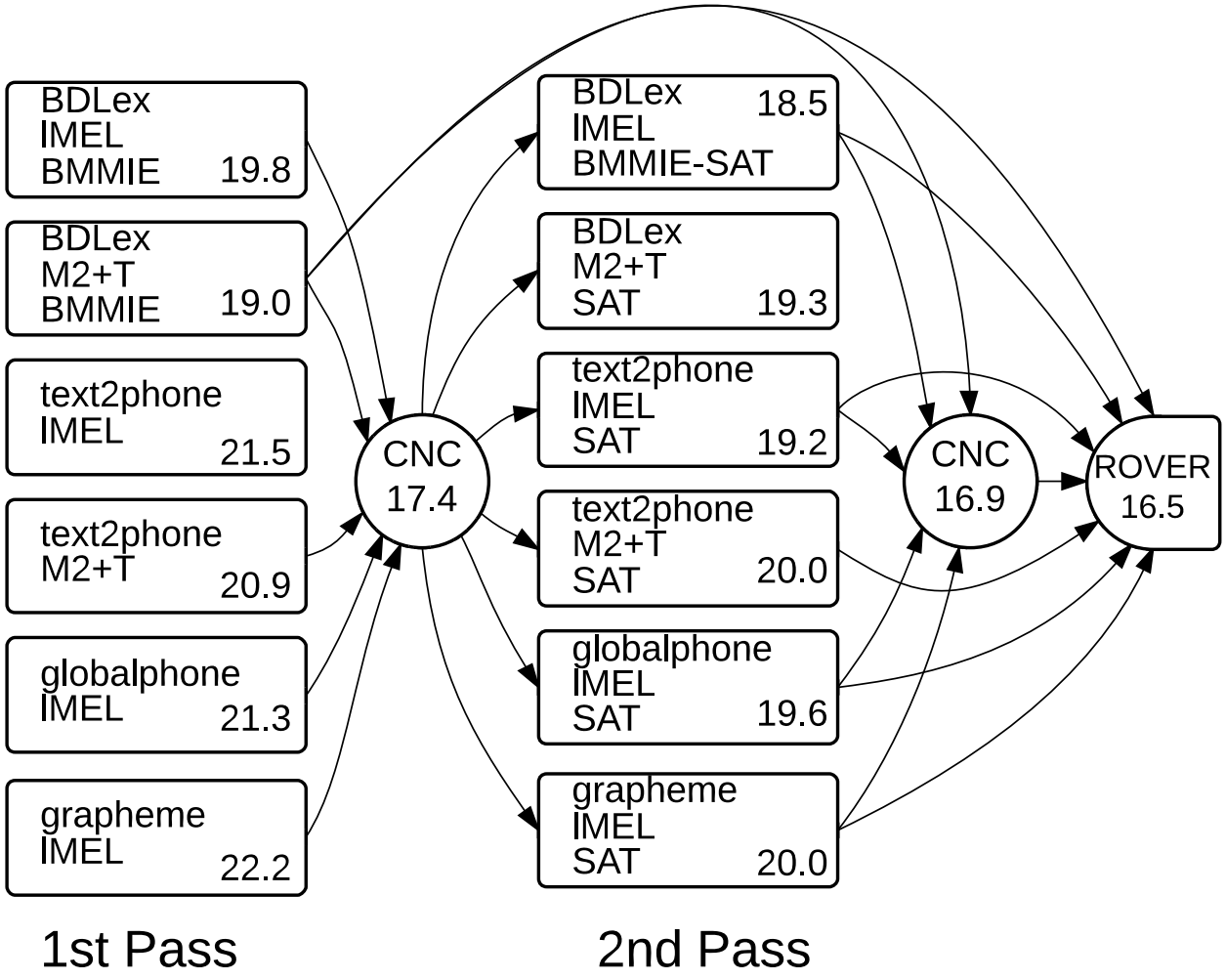
Effect on WER

Training material	WER (CI)
baseline (Quaero core)	27.0
baseline + unfiltered add. Quaero & Ester	29.0
baseline + filtered add. Quaero (75%)	26.9
baseline + filtered add. Quaero (50%)	26.5
baseline + filt'd add. Q. (50%) + Ester 1, 2, & dev (37%)	26.3

Effect on DB Size

Source	Utts. (K)			Hours		
	Unfilt.'d	Kept	%	Unfilt.'d	Kept	%
Quaero core	32.5	32.5	100	140	140	100
add. Quaero + Ester	31.7	18.1	56.9	162	128	78.9
Total	64.2	50.6	78.72	302	267	88.7

System Results on Eval 2012



Conclusion

System Results & Conclusion

(Sub)System	eval2011	eval2012	eval2013
1st Single Best	18.6	19.0	18.6
2nd Pseudograph.	n.a.	20.0	18.9
ROVER	15.6	16.5	15.4

Conclusions

- Created pseudographeme approach for French
- DBNFs & tonal features give significant gains
  - IMEL DBNFs better with SAT than M2+T
- Gains from filtering acoustic training data
- Proper elision normalisation is important

Karlsruhe Institute of Technology

Thank you!