Karlsruhe Institute of Technology

Error Blaming based on Decoding Output

Mark Erhardt, Dominic Telaar, Tanja Schultz
presented by Tim Schlippe

KIT, Institute for Anthropomatics, Cognitive Systems Lab

Overview

- Motivation
- Error Blaming
- BioKIT
- Error Categories and Regions
- Error Blaming Output
- Exemplary Blaming
- Conclusion

Motivation

Problem: Improving ASR systems is arduous.

- Automatic Speech Recognition (ASR) systems are complex
- Not easy to identify room for improvement based on evaluation results alone due to combined effect of all existing errors
- Exploring the complete parameter space is unfeasible.

Approach: Implementation of an Error Blaming tool to BioKIT for …

- automatic analysis of decoding by comparing reference and hypothesis
- dividing utterances into error regions
- assigning error regions to error categories

➔ Goals: Improving ASR systems, improving algorithms and debugging BioKIT

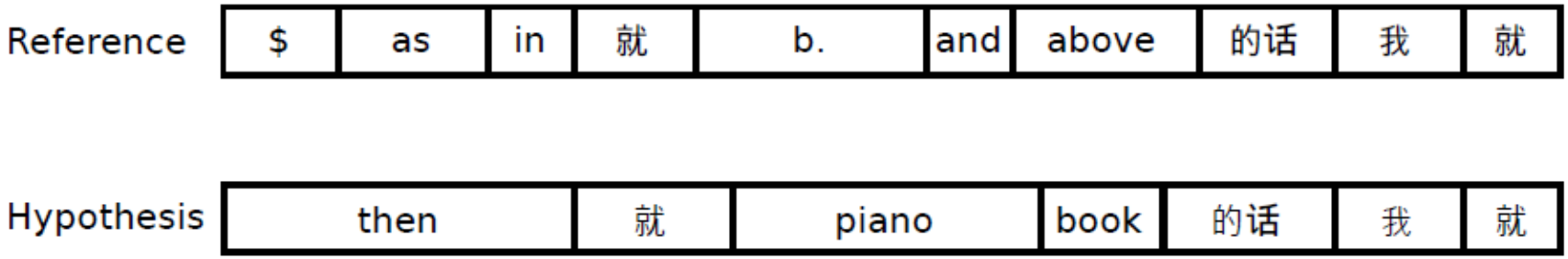
Error Blaming

- Identify problematic regions and causes, and put blame on components, e.g. Acoustic/Language model

  [Chase et al., 1997] [Nanjo et al., 1999]
- Frame-wise analysis of the decoding to estimate impact of different pruning parameters [Steinbiss et al., 2010]

- [Chase et al., 1997]
  Chase, Lin L. *Error-responsive feedback mechanisms for speech recognizers*. Pittsburgh, PA: Carnegie Mellon University, 1997.
- [Nanjo et al., 1999]
  Hiroaki Nanjo, Akinobu Ri, and Tatsuya Kawahara, Automatic Diagnosis of Recognition Errors in Large Vocabulary Continuous Speech Recognition System," Joho Shori Gakkai Kenkyu Hokoku, vol. 99, no. 64, pp. 41–48, 1999.
- [Steinbiss et al., 2010]
  Volker Steinbiss, Martin Sundermeyer, and Hermann Ney, "Direct Observation of Pruning Errors (DOPE): A Search Analysis Tool", I nterspeech, pp. 214–217, 2010.

What is Error Blaming

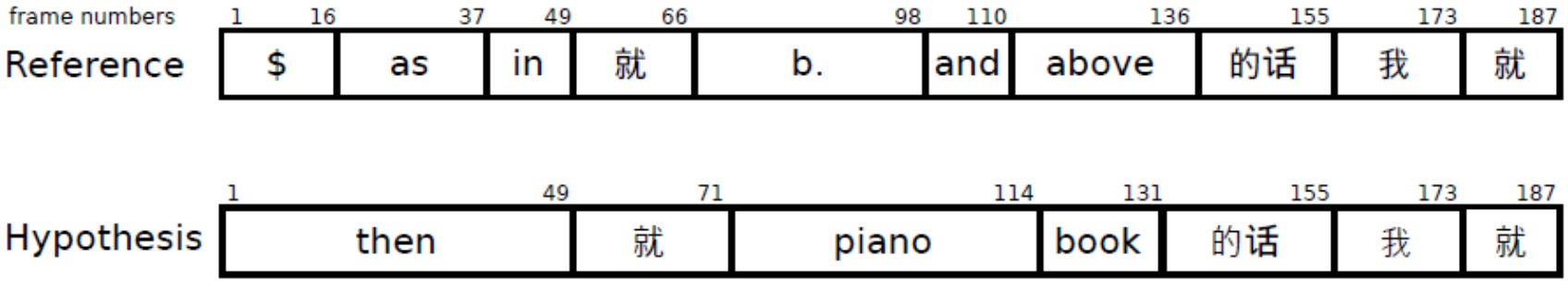


Example from the SEAME Code-Switching corpus [Vu et al., 2012]

- Reference and Hypothesis
  - Reference — Best result from forced alignment.
  - Hypothesis — Best result from unimpeded decoding.

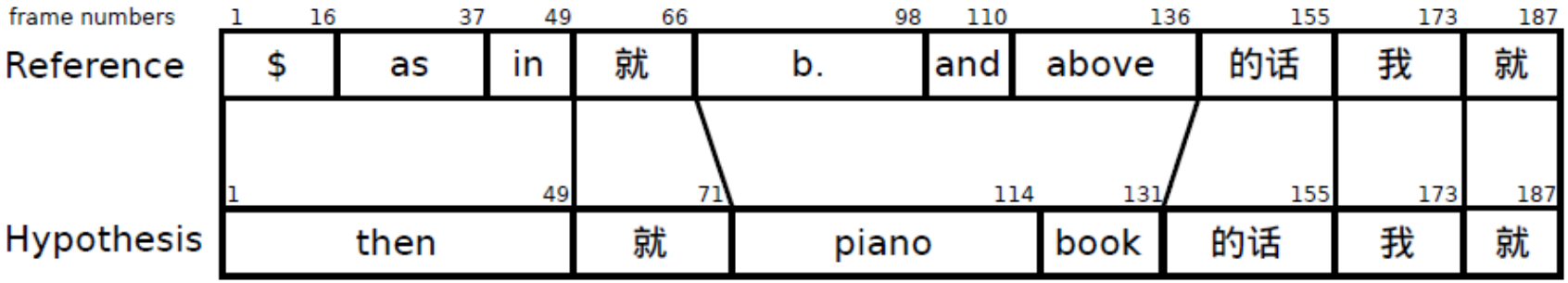
- [Vu et al., 2012]
  Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng Sion Chng and Tanja Schultz, „A first speech recognition system for mandarin-english code-switch conversational speech“, ICASSP 2012, pp. 4889–4892, 2012.

What is Error Blaming



- On the basis of word boundary information, …

What is Error Blaming

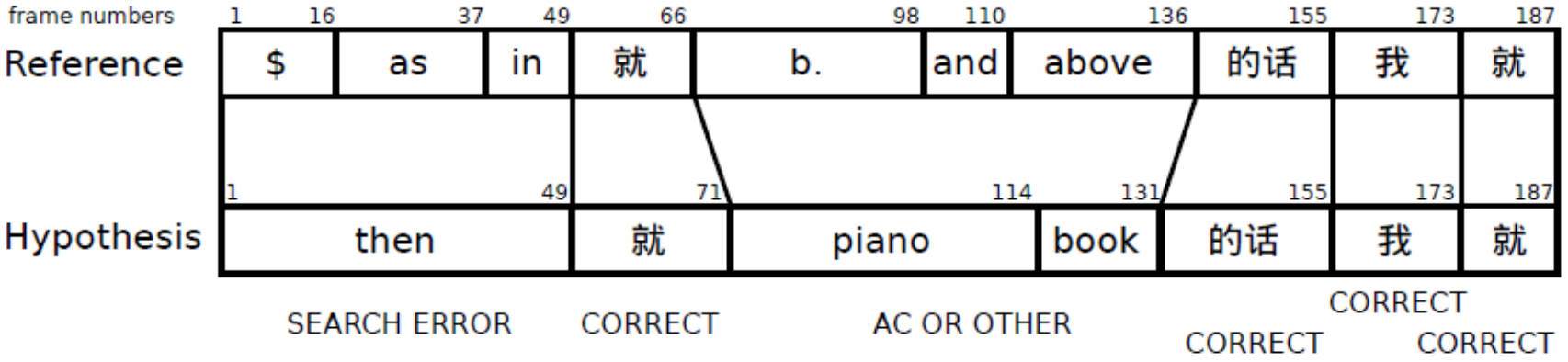


- On the basis of word boundary information,

  we can segment an utterance into error regions.

  - Aligned regions of reference and hypothesis
    - Have similar length and location
    - Separated by regions with correctly recognized words

What is Error Blaming



- On the basis of word boundary information,

  we can segment an utterance into error regions.

- Then assign them to Error Categories.

BioKIT

- Process various signals e.g. Speech, muscle, motion, and brain activity
- Dynamic decoder
- Online-capable
- Integrated Error Blaming
- Two-layered structure
  - Core functionalities in C++
  - Interface to Python

Error Categories

Based on [Chase et al., 1997] and [Nanjo et al., 1999]

- (Correct)
- Search Error → Decoder
- Homophone (e.g. *for*, *four*) → Language Model
- LM overwhelm → Language Model
- AM & LM overwhelm → Acoustic Model and Language Model
- AM or Other → Acoustic Model, Noise, Mispronunciation, Stutter or Transcript Errors

LM = Language Model
AM = Acoustic Model

Blaming Output (Error Region Alignment)

What are Blaming Results?

- Result of Error Blaming Procedure on Decoding Output
- In-detail comparison of Reference and Hypothesis
- Comma-separated value (.csv) table

  - AM and LM Score
  - HMM state sequence
  - Length and location of region
  - Error Category

Reference-Frames:	73 - 85	86 - 103	104 - 147
Hypothesis-Frames:	73 - 85	86 - 103	104 - 147
Reference:	with(3)	that	surge has(2)
Hypothesis:	with(3)	that	searches
AM-Ref:	758.64	995.68	2799.51
AM-Hypo:	758.64	995.68	2725.93
Average AM-Ref:	58.36	55.32	63.63
Average AM-Hypo:	58.36	55.32	61.95
LM-Ref:	5.06	58.48	179.89
LM-Hypo:	5.06	58.48	136.58
Error-Category:	CORRECT	CORRECT	AC_LM_OVERWHELM

Blaming Output (Error Region Alignment)

A Word to Scores

- **AM-Score** Sum of acoustic model scores in current word.
- **LM-Score** Likelihood for word to occur in this context.
- **Total Score** Sum of AM-Score and LM-Score from beginning of utterance to current position.

Please note:

LOWER score is better!

HMM state sequence of traversed AMs not displayed here.

Reference-Frames:	73 - 85	86 - 103	104 - 147
Hypothesis-Frames:	73 - 85	86 - 103	104 - 147
Reference:	with(3)	that	surge has(2)
Hypothesis:	with(3)	that	searches
AM-Ref:	758.64	995.68	2799.51
AM-Hypo:	758.64	995.68	2725.93
Average AM-Ref:	58.36	55.32	63.63
Average AM-Hypo:	58.36	55.32	61.95
LM-Ref:	5.06	58.48	179.89
LM-Hypo:	5.06	58.48	136.58
Error-Category:	CORRECT	CORRECT	AC_LM_OVERWHELM

Blaming Output (Error Region Alignment)

Reference and Hypothesis

- **Reference** Best result from forced alignment.
- **Hypothesis** Best result from unimpeded decoding.

Reference-Frames:	73 - 85	86 - 103	104 - 147
Hypothesis-Frames:	73 - 85	86 - 103	104 - 147
Reference:	with(3)	that	surge has(2)
Hypothesis:	with(3)	that	searches
AM-Ref:	758.64	995.68	2799.51
AM-Hypo:	758.64	995.68	2725.93
Average AM-Ref:	58.36	55.32	63.63
Average AM-Hypo:	58.36	55.32	61.95
LM-Ref:	5.06	58.48	179.89
LM-Hypo:	5.06	58.48	136.58
Error-Category:	CORRECT	CORRECT	AC_LM_OVERWHELM

Blaming Output (Error Region Alignment)

Reference and Hypothesis

- **Reference** Best result from forced alignment.
- **Hypothesis** Best result from unimpeded decoding.

Reference-Frames:	73 - 85	86 - 103	104 - 147
Hypothesis-Frames:	73 - 85	86 - 103	104 - 147
Reference:	with(3)	that	surge has(2)
Hypothesis:	with(3)	that	searches
AM-Ref:	758.64	995.68	2799.51
AM-Hypo:	758.64	995.68	2725.93
Average AM-Ref:	58.36	55.32	63.63
Average AM-Hypo:	58.36	55.32	61.95
LM-Ref:	5.06	58.48	179.89
LM-Hypo:	5.06	58.48	136.58
Error-Category:	CORRECT	CORRECT	AC_LM_OVERWHELM

Blaming Output (Error Region Alignment)

Reference and Hypothesis

- **Reference** Best result from forced alignment.
- **Hypothesis** Best result from unimpeded decoding.

Reference-Frames:	73 - 85	86 - 103	104 - 147
Hypothesis-Frames:	73 - 85	86 - 103	104 - 147
Reference:	with(3)	that	surge has(2)
Hypothesis:	with(3)	that	searches
AM-Ref:	758.64	995.68	2799.51
AM-Hypo:	758.64	995.68	2725.93
Average AM-Ref:	58.36	55.32	63.63
Average AM-Hypo:	58.36	55.32	61.95
LM-Ref:	5.06	58.48	179.89
LM-Hypo:	5.06	58.48	136.58
Error-Category:	CORRECT	CORRECT	AC_LM_OVERWHELM

Blaming Output (Error Region Alignment)

Error Regions

- Segments of utterance
- Comparable content in Reference and Hypothesis

	1st error region	2nd error region	3rd error region
Reference-Frames:	73 - 85	86 - 103	104 - 147
Hypothesis-Frames:	73 - 85	86 - 103	104 - 147
Reference:	with(3)	that	surge has(2)
Hypothesis:	with(3)	that	searches
AM-Ref:	758.64	995.68	2799.51
AM-Hypo:	758.64	995.68	2725.93
Average AM-Ref:	58.36	55.32	63.63
Average AM-Hypo:	58.36	55.32	61.95
LM-Ref:	5.06	58.48	179.89
LM-Hypo:	5.06	58.48	136.58
Error-Category:	CORRECT	CORRECT	AC_LM_OVERWHELM

Blaming Output (Error Region Alignment)

Error Categories

- Hint at responsible error sources

Reference-Frames:	73 - 85	86 - 103	104 - 147
Hypothesis-Frames:	73 - 85	86 - 103	104 - 147
Reference:	with(3)	that	surge has(2)
Hypothesis:	with(3)	that	searches
AM-Ref:	758.64	995.68	2799.51
AM-Hypo:	758.64	995.68	2725.93
Average AM-Ref:	58.36	55.32	63.63
Average AM-Hypo:	58.36	55.32	61.95
LM-Ref:	5.06	58.48	179.89
LM-Hypo:	5.06	58.48	136.58
Error-Category:	CORRECT	CORRECT	AC_LM_OVERWHELM

Better scores in Hypothesis

Blaming Output (Acoustic Model Confusion Table)

Acoustic Model Confusions

- Confusion Pairs of reference and hypothesis acoustic model
  - Number of occurences of confusion pair

  → Impact of confusion pair

Ref. model	Hypo model	Occ.	Mean Ref. Score	Mean Score dist.	Score dist. sdev.	Total Conf. Ref. model	Total Occ. Ref. model
a_ME(|)-m(42)	a_ME(|)-m(17)	27	54.61	6.15	3.74	177	177
o(|)-e(36)	o(|)-e(13)	29	73.07	17.91	9.27	84	84
n_ME(|)-e(14)	n_ME(|)-e(13)	22	57.85	3.06	3.50	42	50

Blaming Output (Acoustic Model Confusion Table)

Acoustic Model Confusions

- Confusion Pairs of reference and hypothesis acoustic model
  - Average AM score in reference

Ref. model	Hypo model	Occ.	Mean Ref. Score	Mean Score dist.	Score dist. sdev.	Total Conf. Ref. model	Total Occ. Ref. model
a_ME(|)-m(42)	a_ME(|)-m(17)	27	54.61	6.15	3.74	177	177
o(|)-e(36)	o(|)-e(13)	29	73.07	17.91	9.27	84	84
n_ME(|)-e(14)	n_ME(|)-e(13)	22	57.85	3.06	3.50	42	50

Blaming Output (Acoustic Model Confusion Table)

Acoustic Model Confusions

- Confusion Pairs of reference and hypothesis acoustic model
  - Mean AM Score Distance of hypothesis to reference (= Ref-AM – Hyp-AM)
  - with standard deviation of AM Mean Score distance

→ e.g. Mean AM Score Distance small → combine model may be helpful

Ref. model	Hypo model	Occ.	Mean Ref. Score	Mean Score dist.	Score dist. sdev.	Total Conf. Ref. model	Total Occ. Ref. model
a_ME(|)-m(42)	a_ME(|)-m(17)	27	54.61	6.15	3.74	177	177
o(|)-e(36)	o(|)-e(13)	29	73.07	17.91	9.27	84	84
n_ME(|)-e(14)	n_ME(|)-e(13)	22	57.85	3.06	3.50	42	50

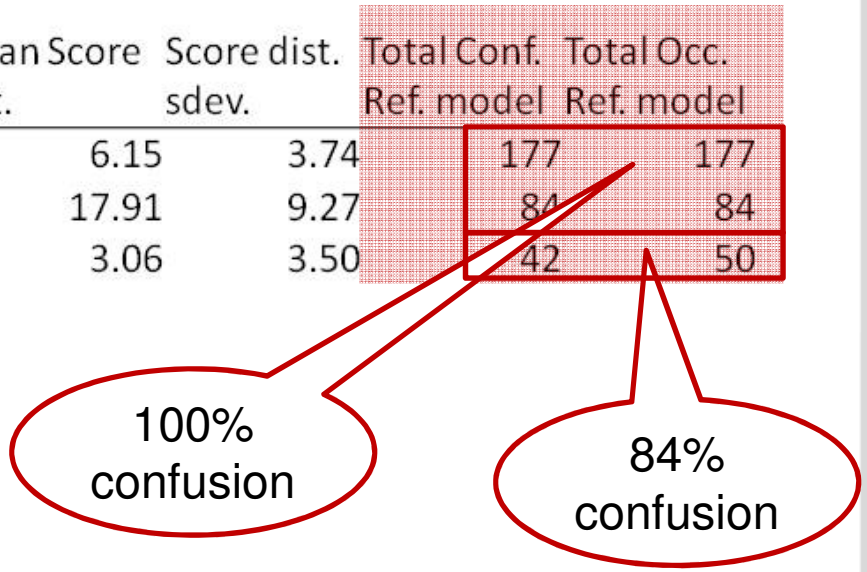
Blaming Output (Acoustic Model Confusion Table)

Acoustic Model Confusions

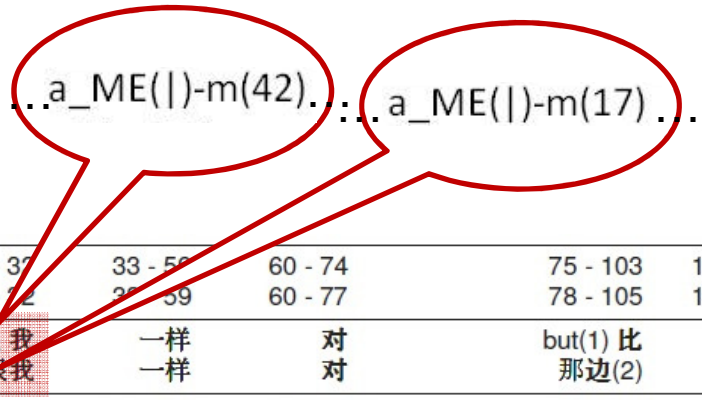
- Confusion Pairs of reference and hypothesis acoustic model
  - Confusions of reference model versus total number of occurences

Ref. model	Hypo model	Occ.	Mean Ref. Score	Mean Score dist.	Score dist. sdev.	Total Conf. Ref. model	Total Occ. Ref. model
a_ME(|)-m(42)	a_ME(|)-m(17)	27	54.61	6.15	3.74	177	177
o(|)-e(36)	o(|)-e(13)	29	73.07	17.91	9.27	84	84
n_ME(|)-e(14)	n_ME(|)-e(13)	22	57.85	3.06	3.50	42	50

Models of the same phoneme at word boundaries and inside words confused



Error Blaming Example



Reference-Frames:	1 - 9	10 - 32	33 - 59	60 - 74	75 - 103	104 - 114	115 - 184
Hypothesis-Frames:	1 - 9	10 - 32	33 - 59	60 - 77	78 - 105	106 - 114	115 - 184
Reference:	$	跟 我	一样	对	but(1) 比	我	小 两 天
Hypothesis:	$	跟我	一样	对	那边(2)	我	想 天
AM-Ref:	390.01	1121.96	1269.22	806.16	1640.61	560.39	3424.45
AM-Hypo:	390.01	1053.39	1269.22	972.46	1614.30	456.88	3433.66
Average AM-Ref:	43.33	48.78	47.01	53.74	56.57	50.94	48.92
Average AM-Hypo:	43.33	45.80	47.01	54.03	57.65	50.76	49.05
LM-Ref:	60.00	147.74	85.72	83.40	182.68	24.66	308.83
LM-Hypo:	60.00	108.28	61.23	80.68	75.44	63.73	185.76
Error-Category:	CORRECT	AC_LM_OVERWHELM	CORRECT	CORRECT	AC_LM_OVERWHELM	CORRECT	LM_OVERWHELM

- Exemplary blaming on SEAME Code-Switching corpus
  - Models at beginning and end of words are substituted (Models with 100% confusion rate)
  - Correct Mandarin characters but differing segmentation

Error Blaming Example

- Exemplary blaming on SEAME Code-Switching corpus
  - Correct Mandarin characters but differing segmentation
  - Models at beginning and end of words are substituted (Models with 100% confusion rate)
  - ➔ Context tree distinguished between phones at word boundaries (due to tags in dictionary)
  - ➔ Caused more than one set of AMs for the same Mandarin sequence
  - ➔ Removed word boundary tags

System	MER Devset	MER Testset
Baseline system	46.9%	36.0%
Improved system	44.9%	31.5%
Absolute Improvement	2.0%	4.5%
Relative Improvement	4.3%	12.5%

Conclusion and Future Work

- Error Blaming for BioKIT through
  - Segmenting utterances into Error Regions
  - Comparing scores of Reference and Hypothesis
  - Assigning Error Regions to Error Categories
- Error Categories give starting point for error investigation
- Model Confusions point out potential problematic acoustic models
- Useful to improve individual recognition systems
- Useful for testing algorithms in decoder
- Future work will include further model dependent analysis of errors

Thanks for your interest!

References

- [1] Lin L Chase, Error-Responsive Feedback Mechanisms for Speech Recognizers, Ph.D. thesis, Pittsburgh, PA: Carnegie Mellon University, 1997.
- [2] Hiroaki Nanjo, Akinobu Ri, and Tatsuya Kawahara, “Automatic Diagnosis of Recognition Errors in Large Vocabulary Continuous Speech Recognition System,” Joho Shori Gakkai Kenkyu Hokoku, vol. 99, no. 64, pp. 41–48, 1999.
- [3] Volker Steinbiss, Martin Sundermeyer, and Hermann Ney, “Direct Observation of Pruning Errors (DOPE): A Search Analysis Tool”, Interspeech, pp. 214–217, 2010.
- [4] Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng Sion Chng and Tanja Schultz, „A first speech recognition system for mandarin-english code-switch conversational speech“, ICASSP, pp. 4889–4892, 2012.