



CRIM'S SPEECH TRANSCRIPTION SYSTEM FOR ETAPE 2011

Vishwa Gupta, Gilles Boulianne, Frederick
Osterrath, Pierre Ouellet

June 19, 2012





Acoustic Model Training Using Kaldi Toolkit

- Training data: 300 hours (Ester) + 178 hours (Quebec French)
- 26 MFCC+delta MFCC X 9 frames \rightarrow LDA \rightarrow 40 features/frame
- Train 6 different models:
 - 1: SI: 300 hours, 5k states, 150k Gaussians, sil penalty, MMI
 - 2: SAT: 478 hours, 5k states, 400k Gaussians, fMMI+BMMI
 - 3: SAT: 478 hours, 10 states, 200k Gaussians, fMMI+BMMI
 - 4: SAT: 300 hours, 10k states, 200k Gaussians, fMMI+BMMI
 - 5: SAT: 300 hrs, 10k states, 200k Gaussians, silence penalty
 - 6: SGMM: trained from model 4, SAT



Typical Recognition Scenario

- Diarize audio file. For each segment do:
- SI decoding (model 1) → decoded lattices
- Decoded lattices → fMLLR transform/speaker for SAT model
- Decoding with SAT model → SAT decoded lattices
- SAT decoded lattices → rescore with quadgram LM
- Rescored lattices → confusion network (MBR decoding)



Typical Word Error Rate (WER) for EVAL Data

- SI recognition (with model 1) → 31.0%
- Decoding with SAT model (model 2) → 27.9%
- Rescore with quadgram LM → 26.4%
- Confusion network (MBR decoding) → 26.03%



Fixed Versus Variable Frame Rate Decoding

With Variable Frame Rate Decoding:

- Dev set: 0.3% absolute WER reduction
- Test set: 0.1% absolute WER reduction
- Test set: 0.5% absolute WER reduction after ROVER



ROVER RESULTS

Step	details	Dev	Eval
9a	Model 2 vfr MBR	24.2	26.03/26.54
9b	Model 3 MBR	24.6	26.4
9c	Model 3 vfr MBR		26.3
9d	Model 4 MBR	24.4	26.4
9e	Model 4 vfr MBR		26.4
9f	Model 5 vfr MBR	25.4	27.0
9g	SGMM vfr MBR	26.5	28.0
10a	9a to 9g		25.26
10b	9a-9c-9e-9f-9g (all vfr)		24.48/25.25
10c	9a-9b-9d-9f-9g (3 vfr, 2 fixed)	23.35	24.93
10d	9a-9f-9g (all vfr)	23.33	24.77



CPU Times

- Single system time is 12.2 times real-time.

Combination	N. of systems	\times RT
9a	1	12.2
10a	7	46.5
10b	5	34.3
10c	5	33.3



Language Model Sources and Perplexity

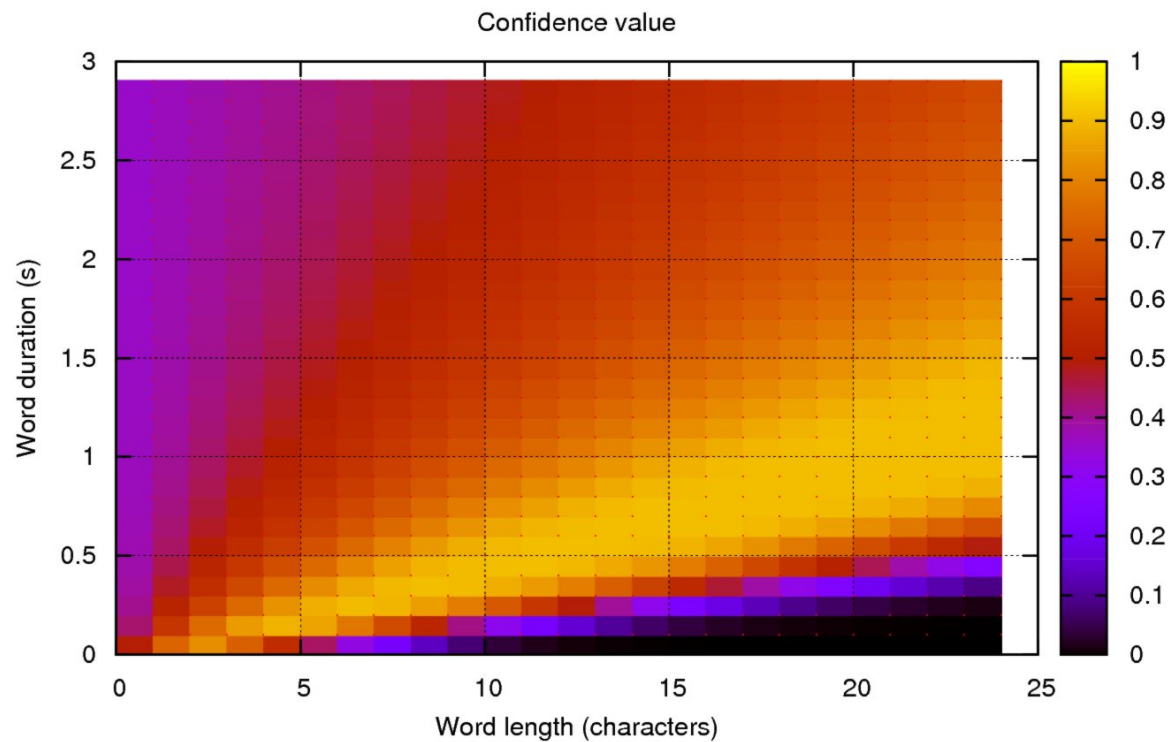
- *Text selection* is from the French gigaword database
- Google 4-grams: from 118 to 109 on development set.

Source	Year	N. words	Weight	PPL
Text selection	1995-2008	8.1 M	0.27	171
ETAPE train	2010-2011	200 K	0.23	227
Le Monde	1987-2003	343 M	0.20	188
Google 4-grams	2008	100 T	0.15	2203
EPAC trans.	2003-2004	1.3 M	0.15	203
ESTER trans.	2003	378 K	0.03	317
Combined				97



Confidence Scoring

- Confidence = $0.75 \times \text{WP} + 0.25 \times \text{DP}$
- DP: asymmetric Gaussian distribution, $G(l/d)$





Vocabulary

- Initial : highest weighted frequency from LM database
- Add 10,000 words from:
 - ETAPE, EPAC, ESTER
 - French Departments, Paris Metro Stations, French Acronyms
- Out-of-vocabulary rate 0.65% on development set



Case Sensitivity

- Vocabulary (or dictionary) is case-sensitive
- LM contains case-sensitive n-grams
- Achieve best results for case-sensitive transcription
- Alternate Scenario:
 - 1) decode with case-independent dictionary/LM
 - 2) rescore final transcript with case-sensitive LM



Conclusions

- Kaldi toolkit instrumental in getting very good results
- Case sensitive LM and dictionary give good case dependent transcription
- Silence model penalty improves alignment in music segments and reduces WER.