Lightly Supervised Acoustic Model Training on EPPS Recordings

Matthias Paulik and Alex Waibel
Interactive Systems Laboratories, Carnegie Mellon University / Universität Karlsruhe (TH)

(1) Introduction

- **European Parliament Plenary Sessions (EPPS)**
  - Broadcast live
  - Proceedings publicly available

  } in n languages

interpreters

L_1: p_1 $i_{2\to1}$ $i_{3\to1}$ . . .
L_2: $i_{1\to2}$ p_2 $i_{3\to2}$ . . .
L_3: $i_{1\to3}$ $i_{2\to3}$ p_3 . . .

n audio channels
interpreters i providing the necessary translations

p_1 p_2 p_3 . . .

politicians p speaking in different official languages L of the EU

live broadcast

off-line

transcribers

Hello good morning I give the floor to Mr. Mueller. Meine Damen und Herren, ich spreche heute

rainbow text edition (RTE)

translators
~ 2 months

FTE

final text edition available in all official languages L

- **Train ASR for language L_i by exploiting available EPPS data**
- **Problem: FTE can differ significantly from politician and interpreter speech**

⇒ *Apply lightly supervised acoustic model training*

- ✗ Bias initial ASR with available knowledge
- ✗ Use hypotheses of biased system for training

- **In this work: bias initial German ASR with**
  - German FTE
  - German translations extracted from English and Spanish audio channels using spoken language translation (SLT)

Comparing politician speech (p) and FTE (f):

(p) But it must be a policy that is shared in partnership with Russia not a covert <hesitation> cover for directing ...

(f) However, it must be a shared policy in partnership with Russia not a covert way of directing ...

Comparing interpreter speech (i), FTE (f) and a human translation of respective Spanish interpreter speech (s):

(i) Mister Poettering President President of the Commission, we confirmed with a great majority the Commission President designate, J. Barroso ...

(f) Mister President of the Commission, J. Barroso was elected by a large majority as the next President of the European Commission ...

(s) Mister President of the Commission, J. Barroso, the future President of the European Commission, was elected by a broad majority ...

(2) Experimental Setup

- **Data: German, English, Spanish EPPS audio**
  - Satellite live broadcast recorded @ interACT

	Sessions	Utterances	Audio [h]
Dev	1	592	1.69
Test	1	885	2.04
Training	93	73,408	142.74

German audio data statistics

  - German dev/test transcriptions created @ interACT
  - No human transcriptions created for Eng, Spa
  - FTEs extracted from Europarl_v3 corpus
- **ASR Systems: German, English, Spanish**
  - 2 pass systems; Janus Recognition Toolkit
  - Ger: AM - 70h BN data; LM - German FTEs
  - Eng, Spa: interACT '06 TC-STAR sub-systems
  - Typical WER: 12-13% Eng and 11-12% Spa
- **MT Systems: English→German, Spanish→German**
  - trained on Europarl_v3 corpus
  - GIZA++, interACT STTK decoder
  - Dev/test BLEU: 12.5/15.2 E→G, 11.9/13.4 S→G

(3) FTE & SLT based Supervision: Impact on WER

- **Session specific supervision via biased LM**
- **3 Types of supervision**
  - **GFTE sup.**: FTE is part of LM training data
  - **FTE sup.**: FTE receives higher interpolation weight
  - **SLT sup.**: include 1000-best SLT hypotheses (per source language) in LM training, use interpolation
- **No supervision for Eng/Spa ASR**
- **GFTE supervision for MT systems (TM and LM)**

		-	GFTE	FTE	FTE & SLT
PPL	Dev	219	206	161	138_e 142_s **130**
	Test	190	176	146	127_e 130_s **118**
WER	Dev	22.3	21.6	20.9	20.7_e 20.3_s **20.1**
	Test	21.0	20.1	19.4	19.1_e 19.2_s **18.8**

German LM perplexity and WER for different types of supervision

➔ **Significant gain in transcription performance**

➔ **Gain for SLT based supervision depends on number of languages used**

(4) Acoustic Model Training: Results

- **Decode training data with biased ASR**
- **Apply simple rule based filter to remove noisy and low confidence utterances**
- **2 iterations of Viterbi training**
- **Add 3rd decoding pass to test new AM**

	-		FTE & SLT	
	Dev	**Test**	**Dev**	**Test**
Original AM	22.2	21.2	20.6	19.2
GFTE AM	20.9	20.0	18.8	18.4
FTE AM	20.7	19.9	18.8	18.2
FTE + SLT AM	20.4	19.8	18.8	18.0

3rd pass WER with different acoustic models and applying either no supervision or FTE & SLT based supervision on dev/test

(5) Conclusion & Future Work

- **14.3% relative improvement in WER**
- **Future Work:**
  - apply SLT based supervision on utterance level
  - include additional languages
  - more sophisticated filtering scheme based on ASR word confidences