Lightly Supervised Acoustic Model Training on EPPS Recordings

Matthias Paulik and Alex Waibel

Interactive Systems Laboratories (interACT)
Carnegie Mellon University (USA), Universität Karlsruhe TH (Germany)
{paulik, waibel}@cs.cmu.edu

Abstract

Debates in the European Parliament are simultaneously translated into the official languages of the Union. These interpretations are broadcast live via satellite on separate audio channels. After several months, the parliamentary proceedings are published as final text editions (FTE). FTEs are formatted for an easy readability and can differ significantly from the original speeches and the live broadcast interpretations. We examine the impact on German word error rate (WER) when introducing supervision based on German FTEs and supervision based on German automatic translations extracted from the English and Spanish audio. We show that FTE based supervision and additional interpretation based supervision provide significant reductions in WER. We successfully apply FTE supervised acoustic model (AM) training using 143h of recordings. Combining the new AM with the mentioned supervision techniques, we achieve a significant WER reduction of 13.3% relative.
Index Terms: lightly supervised acoustic model training, EPPS, speech recognition

1. Introduction

European Parliament Plenary Sessions (EPPS) are broadcast live in the official languages of the Union. For each language, an interpreter provides the simultaneous translation (interpretation) whenever a politician is speaking in a language different from the respective target language of the individual interpreter. The proceedings in the parliament are also made available in form of so-called final text editions (FTE). These texts are created by a multitude of human transcribers and translators and are available through the EUROPARL web site [1] in all official languages after approximately two months [2]. FTEs are formatted for an easy readability and can differ significantly from the politicians' speeches and their respective live broadcast interpretations. Figure 1 gives an overview on the overall process and also shows examples for the differences between the original speeches, their final text editions and the broadcast interpretations.

Unsupervised AM training is based on speech data for which no human transcriptions are available. Training relies in that case on automatic transcriptions that are created with an initial ASR system. Lightly supervised AM training [3] refers to the case where some imperfect human transcriptions, for example closed-captions provided during television broadcasts, can be used to either bias the initial ASR system for an improved transcription performance or to filter erroneous ASR hypotheses. Given the availability of final text edition and several audio channels in different languages, recordings of EPPS are suitable for lightly supervised acoustic model (AM) training. Supervision for AM training in language L_i

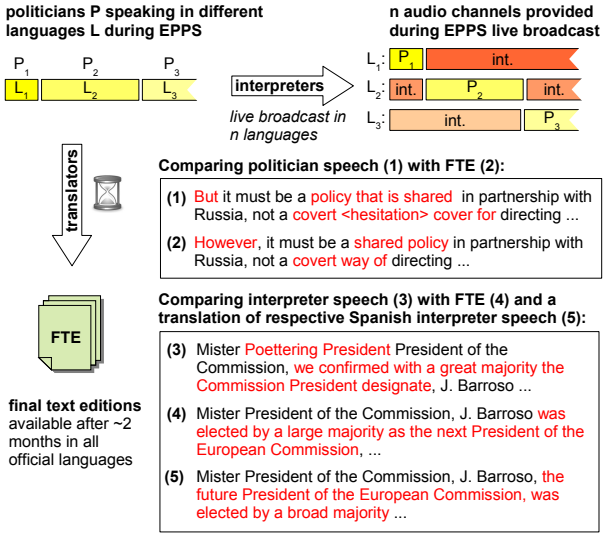

Figure 1: Overview on the available knowledge source to introduce supervision for ASR on EPPS recordings .

can be introduced via the respective FTE in language L_i or via automatic translations into language L_i from FTEs and interpretations available in languages $L_{j \neq i}$. Gollan et al. mention in [2] the possibility to include English FTEs in the language model (LM) training data to achieve a lightly supervised AM training on English EPPS recordings. However, no experiments for FTE based supervision are reported in [2] since the necessary FTEs were not available at that time.

In our work, we examine the impact of FTE based and interpretation based supervision on German word error rate (WER). We make use of the German FTEs and of German automatic translations extracted from the English and Spanish audio channels. Supervision is introduced via language model interpolation on a per session basis. We show that FTE based supervision as well as a combination of FTE based and interpretation based supervision provides significant improvements in transcription performance and we show that the gain in performance for interpretation based supervision depends on the number of interpretation languages used. We also give results for AM training with automatic transcriptions that were created using FTE based supervision.

The remainder of the paper is organized as follows. Special issues regarding the nature of FTEs and interpretations are discussed in Section 2. Section 3 gives a brief description of the used data and our ASR and machine translation (MT) sys-

	sessions	utterances	audio [h]
dev	1	592	1.69
test	1	885	2.04
training	93	73,408	142.74

Table 1: Data statistics.

tems. Experiments and results regarding the impact on transcription performance when applying FTE based supervision and English/Spanish interpretation based supervision are given in 4. Finally, in Section 5 we describe the experiments for FTE supervised AM training.

2. The Nature of Interpretation and FTE

Interpreters apply special strategies to keep pace with the source speaker. These strategies include planned omissions and synopses. Another strategy is the anticipation of a final verb or syntactic construction before the source speaker has uttered the corresponding constituent [4]. In addition to these strategies, interpreters do also elaborate and change information and they convey not only all elements of meaning, but also the intentions and feelings of the source language speaker [5]. For these reasons, interpretation differs significantly from translation. In this context we can describe the final text edition of an EPPS as a 'formal' style translation of the original speeches given by the politicians. We use the term 'formal' to express the fact that the FTE is not a verbatim transcription or 'literal' translation but rather a text that aims for an easy readability. This means in particular that even the FTE in language L_i with i equal to the language used by the politician can differ significantly from the verbatim transcription of the politician's speech. Examples that show some of the differences between FTE, interpretation and verbatim transcription are given in Figure 1.

3. Experimental Setup

3.1. Data

Table 1 gives an overview on the German audio data statistics of the used development set, test set and non-transcribed training set. All sessions included in the respective sets were recorded in our laboratory from the EPPS satellite live broadcasts in several languages, including German, English and Spanish. The speakers in the recordings switch between interpreters and native or non-native speakers with different, and partly strongly pronounced, accents. Whenever a politician speaks in a language different from the language of the respective audio channel, the microphone is switched back to the interpreter. Delays when the microphone is switched result in short periods of foreign language speech. The German verbatim transcriptions used in this work to measure transcription and translation performance were created in our laboratory. We did not create any English/Spanish reference transcriptions or any additional English/Spanish-to-German reference translations.

3.2. ASR Systems

The employed ASR systems were developed with the Janus Recognition Toolkit (JRTk) [6]. The SRI Language Model Toolkit [7] was used for language model training.

The **German ASR** system features a speaker-independent and a speaker-dependent decoding pass. The ASR subsystems used in the two passes feature a Minimum Variance Distortionless Response (MVDR) front-end [8]. In the first pass a single ASR system is employed to provide the second pass with first best hypotheses for unsupervised speaker adaptation. We apply Maximum Likelihood Linear Regression (MLLR), feature space constrained MLLR (fMLLR) and Vocal Tract Length Normalization (VTLN) for speaker adaptation. The second pass uses two ASR subsystems with slightly different phones sets. At the end of the second pass, confusion network combination (CNC) [9] is applied. The acoustic models were trained on 70h of German broadcast news data. The dictionary consists of 89.6k pronunciation entries. Compound word splitting was employed to keep the out-of-vocabulary rate small. The 4-gram LM was trained on the German final text editions extracted from the Europarl v3 corpus [10]. The LM perplexity and WER for the development set and test set are shown in the first column of Table 2.

The **English ASR** system consists of ASR subsystems [11] that were developed within our laboratory for the TC-STAR [12] Spring 2006 ASR Evaluation. The decoding setup features a speaker-independent pass with two ASR subsystem employing two different acoustic front-ends, namely a traditional Mel-frequency scaled Cepstral Coefficients (MFCC) front-end and a MVDR front-end. CNC is applied at the end of the first pass. The subsequent speaker-dependent pass is performed with only one MFCC based ASR system. The acoustic models were trained on 80h of English EPPS. The dictionary consists of 47K pronunciation entries. The 4-gram LM was trained on the in 2006 within TC-STAR available EPPS transcriptions and final text editions, the Hub4 Broadcast News data and the UN Parallel Text Corpus v1.0.

The decoding setup for the **Spanish ASR** is identical to the setup of the English ASR system. The acoustic models were trained on 140h of Spanish EPPS and Spanish Parliament (CORTES) data. The case-sensitive pronunciation dictionary has 77.9K entries. The 4-gram LM was trained on the within TC-STAR available Spanish EPPS and CORTES texts and transcriptions. We lowercase the Spanish ASR output before handing it over to the MT systems.

The English and Spanish WER on the used development set and test set could not be computed since we did not have the necessary English and Spanish verbatim transcriptions available. The typical word error rate on this task when applying the described decoding setup ranges from 11% to 12% for the Spanish ASR system and from 12% to 13% for the English ASR system.

3.3. MT Systems

The **English-to-German** and **Spanish-to-German** MT systems were trained on the respective parallel parts of the Europarl v3 corpus. Phrase tables were estimated via the GIZA++ toolkit [13] and University of Edinburghs phrase model training scripts. Both systems apply the same 4-gram language model as used in the German ASR system. The STTK beam search decoder [14] combines the translation model scores from the phrase table and the LM scores together with scores from an internal word re-ordering model and simple word and phrase count models to find the best translation. The internal word reordering model assigns higher costs to longer distance reordering. In our experiments we use a reordering

window of 2. The word count model is used to compensate for the tendency of the language model to prefer shorter translations, while the phrase count model can give preference to longer phrases, potentially improving fluency.

The English-to-German and Spanish-to-German automatic translations are computed with the lowercased first best English/Spanish ASR hypotheses as input. The translation reference for BLEU score computation is equal to the German human verbatim transcription used for computing the German ASR word error rate. The BLEU score on the development set is 12.5 for English-to-German and 11.9 for Spanish-to-German. On the test set, the BLEU score is 15.2 and 13.4, respectively. Although the Spanish ASR word error rate is typically approximately 1% lower than the WER of the English ASR, we see a consistently better translation performance for English-to-German translation. This can be explained by the fact that Spanish is a morphologically rich language compared to English.

4. FTE & Interpretation based Supervision: Impact on WER

4.1. Types of Supervision and their Employment

We employ two different types of supervision that are based on the final text editions. In the case where the FTE of a specific session was part of the overall language model training data, we speak of a 'general' FTE (gFTE) supervision. Whenever we mention FTE supervision without the addition of the word 'general' we refer to the case where the final text edition of a specific session receives a higher weight than the remaining LM training data. We achieve this by building a language model on all available FTEs and an LM on the FTE of the respective session. We then interpolate both language models with a fixed interpolation weight of 0.28 for the smaller, session specific language model. This interpolation weight was determined to yield the lowest perplexity on the development set.

In order to introduce interpretation based supervision, we automatically transcribe and translate the English and Spanish audio channels of a session. Using the 1000-best translation hypotheses from each MT system, we build two separate language models and interpolate these. The LM based on the English-to-German translations receives an interpolation weight of 0.54. Finally, we interpolate this LM with the FTE supervised LM, where the interpolation weight for the translation based LM is set to 0.34. The used interpolation weights were again determined on the development set to yield a minimal perplexity. In this context it should be mentioned that, when using the English and Spanish interpretations, it is directly possible to build not only session specific language models, but also utterance specific LMs, as we have shown for example in [15, 16]. This is possible since the interpreters provide the simultaneous translation of the politician's speech with only a minimal delay. Such a more fine grained supervision usually yields a higher performance. However, in this work we only apply supervision that is specific to the individual sessions.

The English and Spanish ASR systems used to transcribe the English and Spanish audio channels do not employ any form of supervision on the development and test set. However, general FTE supervision may be given on parts of the training set. The MT system applies general FTE supervision on development, test and training set, since the according final text editions were part of the translation model and language model training data. No session specific FTE supervision is applied in the MT systems or the English/Spanish ASR systems.

		-	gFTE	FTE	FTE & Int.
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**

Table 2: LM perplexity (PPL) and word error rate for different types of supervision.

4.2. Results

Table 2 shows the German language model perplexity and German word error rate for the different types of supervision. The first column gives the results when no supervision of any form is applied. In the last column we list the results for combining FTE supervision with interpretation based supervision using both, English and Spanish interpretations, as well as the results when using only the English or Spanish interpretation on top of FTE bases supervision. The results show that a significant gain in transcription performance can be achieved by applying FTE supervision and a combination of FTE supervision and interpretation based supervision. The gain in transcription performance for interpretation based supervision depends on the number of interpretation languages used. Complementary information is added with each additional interpretation language.

5. FTE Supervised Acoustic Model Training

We automatically transcribed 142.7h of German EPPS recordings using general FTE supervision and session specific FTE supervision. Before training on these automatic transcriptions, we applied a simple rule based filter to remove noisy and low confidence utterances. The rules for this filter were hand written and tuned on the development set in regards to word error rate. We removed all utterances that had a filler to word ratio that was greater than 2.5 or that had an average word confidence lower than 0.4. During training we did not apply the available word confidence scores in any way. After applying the rule based filter, 57,319 utterances amounting to 137.0h remained in the general FTE supervised case and 56,813 utterances amounting to 136.8h of training data remained in the FTE supervised case. Acoustic model training consisted of two iterations of Viterbi training starting from the original AM.

To test the new acoustic models, we added a simplified third decoding pass to the German decoding setup. This simplified pass features only one ASR system and does not include confusion network combination. Unsupervised speaker adaptation is performed on the CNC output from the second pass. Results are listed in Table 3. It should be noticed that the word error rate for the original acoustic model is worse than the WER achieved after CNC in the second pass. The FTE trained AM outperforms the original AM by 1.0% absolute and the general FTE AM by 0.2% absolute. As shown in Table 2 , the gain in transcription performance when applying FTE supervision instead of gFTE supervision is approximately equal to the gain when introducing interpretation based supervision on top of FTE based supervision. We therefore expect the improvement for acoustic

	dev	test
original AM	20.6	19.2
gFTE trained AM	18.8	18.4
FTE trained AM	18.8	18.2

Table 3: WER before and after FTE supervised acoustic model training.

model training when adding interpretation based supervision to be similar to the improvement seen when applying session specific FTE supervision instead of general FTE supervision. A further improvement can be expected when applying interpretation based supervision on a per utterance basis [15, 16] rather than on a per session basis.

6. Conclusion and Future Work

We examined different available knowledge sources to introduce supervision for automatic speech recognition that is applied to recordings of the German audio channel of European Parliament Plenary Session live broadcasts. These knowledge sources include the German final text editions that are available through the EUROPARL web site approximately two months after the live broadcast. The other used knowledge sources are the English and Spanish interpretations recorded from the respective English/Spanish audio channels of the live broadcasts. We applied English-to-German and Spanish-to-German spoken language translation in order to make use of these interpretations. Supervision was accomplished by biasing the ASR language model on a per session basis via language model interpolation. Starting with a baseline WER of 21.0%, FTE based supervision provided us with an absolute word error rate reduction of 1.6%. By adding interpretation based supervision, we achieved an additional reduction of 0.6% absolute. We showed that the gain in transcription performance for interpretation based supervision depends on the number of interpretation languages used. We successfully applied FTE supervised acoustic model training on 142.7h of German EPPS recordings. The new acoustic model yielded a 1% absolute lower WER compared to the original AM when applying both on a simplified additional decoding pass. The final WER after this additional decoding pass with the new acoustic models was 18.2%. The improvement compared to the baseline WER amounts therefore to 13.3% relative.

We are currently in the progress of automatically transcribing and translating the English and Spanish interpretations of the available training data in order to iteratively perform interpretation supervised acoustic model training. We plan to apply interpretation based supervision for all involved ASR and MT systems on an utterance level as described in [15, 16], rather than on a session level as described in this work. We are also considering to exploit additional interpretation languages and to introduce a more sophisticated filtering scheme based on word confidences to remove erroneous ASR hypotheses from the training data.

7. Acknowledgements

The authors would like to thank Susanne Burger and her team of transcribers for providing the development set and test set German verbatim transcriptions. Thanks also go to Matthias Wölfel and Christian Fügen for providing the baseline German acoustic models and to Florian Kraft for providing the German compound word splitter. The English and Spanish ASR systems used in this work were developed in our laboratory under the integrated project TC-Star -Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org). This work was partly funded by the National Science Foundation (NSF) under the project STR-DUST (Grant number IIS-0325905).

8. References

[1] "EUROPARL: European Parliament web site", http://www.europarl.europa.eu.

[2] Gollan, C., Bisani, M., Kanthak, S. , Schlüter, R. and Ney, H., "Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus", in IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol. 1, pp. 825–828, Philadelphia, USA, Mar 2005.

[3] Lamel, L., Gauvain, J.L. , Adda, G., "Investigating lightly supervised acoustic model training", in Proc. of ICASSP, pp. 477–480, Salt Lake City, USA, May 2001.

[4] Van Besien, F., "Anticipation in Simultaneous Interpretation", in Meta 44(2), pp. 250–259.

[5] Kohn, K. and Kalina, S., "The strategic dimension of interpreting", in Meta 41(1), pp. 118–138.

[6] Soltau, H., Metze, F., Fügen, C. and Waibel, A., "A one pass-decoder based on polymorphic linguistic context assignment", in ASRU, Madonna di Campiglio Trento, Italy, Dec 2001.

[7] Stolcke, A., "SRILM – An Extensible Language Modeling Toolkit", in Intl. Conf. on Spoken Language Processing, Denver, USA, Sep 2002.

[8] Wölfel, M. and McDonough, J., "Minimum Variance Distortionless Response Spectral Estimation", in IEEE Signal Processing Magazine, vol. 22, pp.117–126, Sep 2005.

[9] Mangu, L., Brill, E. and Stolcke, A., "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks", in Computer Speech and Language, pp. 373–400, Apr 2000.

[10] Koehn, P., "Europarl: A Parallel Corpus for Statistical Machine Translation", in MT Summit, Phuket, Thailand, Sep 2005.

[11] Stüker, S., Fügen, C., Hsiao, R., Ikbal, S., Jin, Q., Kraft, F., Paulik, M., Raab, M., Tam, Y., Wölfel, M., "The ISL TC-STAR Spring 2006 ASR Evaluation Systems", in Proc. of the TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, Spain, 2006.

[12] "TC-STAR: Technology and Corpora for Speech to Speech Translation", http://www.tc-star.org.

[13] Och, F.J. and Ney, H., "Improved Statistical Alignment Models", in Proc. of ACL, pp. 440–447, Hongkong, China, 2000.

[14] Vogel, S., "SMT Decoder Dissected: Word Reordering", in International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 2003.

[15] Paulik, M., Stüker, S., Fügen, C., Schultz, T., Schaaf, T. and Waibel, A. "Speech Translation Enhanced Automatic Speech Recognition", in Proc. of ASRU, San Juan, Puerto Rico, 2005.

[16] Paulik, M. and Waibel, A. "Extracting Clues from Human Interpreter Speech for Spoken Language Translation", in Proc. of ICASSP, Las Vegas, USA, 2008.