Adapting Automatic Speech Recognition for Foreign Language Learners in a Serious Game

Joshua Winebarger, Sebastian Stüker, and Alexander Waibel
Insitute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
Karlsruhe, Germany

Abstract

Eveil3d is a project for development of an immersive, virtual "serious game" for computer assisted foreign language learning, with which users interact verbally by means of an automatic speech recogniser. The speech of the target user group, namely adolescent low-proficiency non-native speakers, differs significantly from native adult speakers on which ASR systems are typically trained and thus on which they perform the best. As only a very small corpus of near-domain speech and text is available, the aforementioned difference becomes a development challenge. We deal with this challenge by adapting existing systems on the small data set. We adapt our language models using text selection to augment the in-domain data with similar data from out-of-domain sources. We adapt our acoustic models with MAP adaptation. Through these steps we achieve significant reductions in error.

1 Introduction

Eveil3d is a Franco-German project for the development of a "serious game" for computer assisted language learning of German and French as a foreign language by middle school students in France and Germany, respectively. With the help of virtual reality and head-tracking equipment, students travel to a representation of the Strasbourg Cathedral where they must navigate the environment, solve puzzles and interact with game characters in the target language. Verbal interaction is provided by an automatic speech recognition (ASR) engine.

ASR holds great promise for computer assisted language learning, since it could enable the improvement of oral proficiency, a skill which is one of the most difficult to practice in the classroom (Van Doremalen, Cucchiarini, and Strik 2010). However, the speech produced by early language learners and that of non-adults is difficult to recognise in part due to what is called "mismatch," arising from the so-called "fragility" of ASR. While ASR robustness and the expansiveness of its capabilities have improved dramatically in the last decades, compared to human robustness its performance can still degrade rapidly when testing conditions and data differ from those of training.

The speech of children and to a lesser degree adolescents differs in important ways from that of adults and exhibits significant variability due to developmental changes including anatomical and morphological changes, and greater spectral variability, among others (Potamianos and Narayanan 2003), Changes in vocal tract geometry cause variable effects on fundamental and resonant frequency (Kent 1976). Recognition of low-proficiency non-native speech is also problematic as it deviates from standard in all areas of speech production including morphology, pronunciation, syntax, vocabulary, and sentence structure (Van Doremalen, Cucchiarini, and Strik 2010). As most ASRs are trained for adult, native speech, of which there is also the most training material, such systems make for poor recognisers in our domain (Lawson, Harris, and Grieco 2003; Morgan 2004).

A straightforward solution to mismatch is to train on in-domain data. Strictly speaking, our in-domain data would be in-game speech by French and German adolescent language learners of German and French. Instead, we have a limited amount of near-domain read speech and dialog texts of which there is too little to use for outright training. This combination of unique domain, speech stlye, and limited availbilty of data poses a novel challenge.

Due to the limited amount of data, we attempt instead to adapt existing systems on this data. In this paper we will describe two series of adaptation experiments which we performed for the French language recogniser. First, we adapt the acoustic model (AM) to the near-domain speech by incorporating it into the last step of AM training with a higher weight. We systematically try weight combinations and examine trade-offs between two subsets of near-domain data. For language model (LM) adaptation we try a combination of incorporating invented text and intelligently selecting text from out-of-domain (OOD) corpora, integrating both in our LM.

2 ASR System Description

We develop the ASR systems for Eveil3d using the Janus Recognition Toolkit (JRTk). Front-end preprocessing of the audio sampled at 16kHz produces feature vectors consisting of 13 Mel-Frequency Cepstral Coefficients stacked with a left and right context of seven features. A Linear Discriminant Analysis is then applied to the stacked features, reducing the dimensionality to 40 coefficients.

All acoustic models are based on HMMs with generalized

quinphone states with three states per phoneme and a left-to-right topology without skip states. We used 3000 generalized quinphones found by clustering them with the training data using a decision-tree. Models were trained using incremental splitting of Gaussians. For all models we then estimated one global semi-tied covariance matrix after LDA and refined the models with two iterations of viterbi training.

Language models are 4-gram case-insensitive LMs with either modified Kneser-Ney or Witten-Bell (WB) smoothing. LMs were built using the SRI Language modeling Toolkit (Stolcke 2002). The final LM is a weighted mixture of the component LMs trained on the individual sources, the weights being those minimizing perplexity on our tuning set (a subset of the game dialogs) as found by an expectation-maximization (EM) algorithm in the SRILM Toolkit. We trained a series of case-insensitive LMs from a subset of OOD sources selected automatically to be most similar to the game dialogs using the mixture weighting process mentioned above. These sources are detailed in table 3.

The best-performing LM in the series, using WB smoothing, serves as the baseline LM for the experiments described in section 3. We selected a search vocabulary for our system from the aforementioned sources tuned on a subset of the game dialogs according to the maximum likelihood count estimation described in (Venkataraman and Wang 2003).

3 Adaptation Experiments

3.1 Acoustic Model Adaptation

Acoustic Corpora Our near-domain acoustic data for adaptation and testing comes from three sets of recordings. In all sets, speakers read aloud sentences from the game dialogs as well as from copora from the Quaero project (http://www.quaero.org), automatically selected for their similarity to the game dialogs. The first two sets were recordings of German middle school learners of French at the *A2* level of language proficiency. Speaker ages ranged from 12 to 15 (median 13.) In the first set, which we call *TEST*, 16 speakers spoke 15 utterances each, totaling 20 minutes of speech. This set was used for detecting overfitting. In the second set, which we call *Group A*, 11 speakers read 100 to 150 sentences, or approximately 87 minutes. The third set came from recordings of university students of French of median age 21 reading between 100 and 200 sentences. We selected the six youngest speakers with German as a native language for inclusion in this set, which we called *Group B*. This made up 52 minutes of speech. All speech was then orthographically transcribed in one pass.

We created our acoustic adaptation and our development (*DEV*) sets from combinations of Groups *A* and *B*. The breakdown by group, gender, as well as total utterances and duration in minutes, is shown in Table 1.

Set	*Group A*		*Group B*		Utts	Mins
	♀	♂	♀	♂		
Acoustic adapt.	5	2	2	2	1459	95
DEV	1	3	2	0	609	30

Table 1: Speaker composition and set size of *aADAPT* and *DEV*.

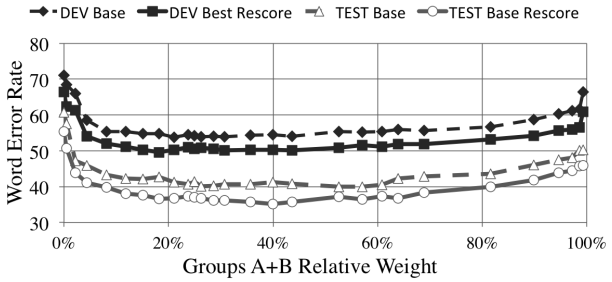

Figure 1: Results of MAP holding ratios of weight factors between *A* and *B* constant. WER on *DEV* and *TEST*

In addition, we dispose of a large amount of out-of-domain broadcast news and conversation domain speech data from the Quaero and Ester (Galliano, Gravier, and Chaubard 2009) projects. This data (*Q+EST*,) is approximately 302 hours long with about 64 thousand utterances.

MAP Adaptation We trained our baseline system on the standard set of Quaero and Ester data. This included context-independent modeling, polyphone clustering, and context-dependent GMM training. The resulting model we then adapt using Maximum A-Posteriori (MAP) adaptation. We perform the MAP adaptation by weighting the Viterbi training statistics from the three sources *A*, *B*, or *Q+EST* with three different weight factors. The effective share of training material or "relative weight" for each source is $R_i = w_i L_i / (\sum_{j \in J} w_j L_j)$ where i is the source in consideration, J is the set of all sources, w_i is the weight factor for source i, and L_i is the length of source i. In the first round of weight adjustment we keep the ratio of the weight factors of *A* and *B* fixed, the factor of *B* being $3/4$ of that of *A*. We chose this factor based on the intuition that the speech in *B* is less useful, as it comes from older and more proficient speakers than that of the targeted game user. We then searched a series of weighting factors for *A+B*, weighting *Q+EST* with a constant factor of 1. Our results are shown in figure 1, where for both *DEV* and *TEST* we show the case-insensitive word error rate (WER) both without lattice rescoring ("Base,") and with the best-performing lattice rescoring ("Best Rescore.") A relative weight of 30.7% to *A+B* gave the lowest WER on *DEV* for all factors tied.

In the second round of experiments, we step through values of factors for *A*, with the factor of *B* being a function of that of *A* such that the relative weight of *A+B* remains at a constant 30.7%. The best combination found that way gave equal factors of 89 for both *A* and *B*, meaning relative weights of 17.6% and 12.4% respectively.

Table 2 summarizes the results. The first line shows the baseline system which does not see any near-domain data, and on which one iteration of Viterbi training is performed on *Q+EST*. The best results on *DEV* and *TEST* from the second round of weighting experiments are shown in line 3, where the WER was reduced by 24-35% relative. Differences in WER within a column are statistically significant within a 95% confidence interval according to the Sign test.

Relative weight R_i (%)			*DEV* WER (CIns)		*TEST* WER (CIns)	
A	*B*	*Q+EST*	Base	Best Rescore	Base	Best Rescore
0	0	100	71.1	66.4	60.8	55.4
0	30.7	69.3	60.1	56.5	48.2	44.5
17.6	13.1	69.3	**54.0**	**50.0**	**39.2**	**35.8**
30.7	0	69.3	55.2	53.4	43.3	40.2

Table 2: Summary of AM adaptation experiments. Best results indicated in boldface.

Source	**Running Words**
l'Humanité (newspaper)	752M
Quaero blog sources	62M
Huffington Post (online newspaper)	5M
Quaero transcripts	2.6M
ESLO (oral corpus) (*http://http://eslo.huma-num.fr/*)	1.1M
CFPP2000 (oral corpus) (*http://ed268.univ-paris3.fr/CFPP2000/*)	417K
TCOF (oral corpus) (*http://www.cnrtl.fr/corpus/tcof/*)	153K

Table 3: Out-of-domain text sources

4 Language Model Adaptation

For language modeling we faced a similar situation as with acoustic modeling. We had a number of out-of-domain (OOD) sources, as well as some 752 lines of game dialogs with 4566 running words. The latter we call *GD*. We automatically selected those texts most similar to the game dialogs using the EM method described in section 2. To expand our relevant data, we had French native speakers and advanced students of French invent in-game utterances, resulting in 1422 additional lines (called *IGD*.) Next, we adapted our LMs using a popular text selection technique described in (Moore and Lewis 2010). The approach uses domain-specific and non-domain specific LMs to compute cross-entropy scores for the out-of-domain data.

First, we concatenated our out-of-domain sources into one set and scored the sentences in the set using the aforementioned approach. We selected the top-scoring percent P of sentences as a new text and included this new text as a mixture source in our LM training data. We concatenated our *GD* and *IGD*, shuffled it, and split it three ways in a cross-validation (CV) approach for determining the best mixture weights on a heldout set, each part being alternately a mixture tuning source, an LM training source, and an in-domain set for text selection. With the best mixture weights determined, we compute the LM with the full *GD+IGD* as LM source and text-selection in-domain set. As a reference, we also computed scores for several LMs without text selection.

We computed perplexity on a test set consisting of the set of sentences read by speakers in the *TEST* set of the previous section. Having optimized our acoustic adaptation, we took the best-performing recogniser from the previous section and tested it with our LMs on both *DEV* and *TEST*. Our results are given in table 4. Those WER with a $\star$ were tested for significance with the Sign test. All were found to be significant.

We achieve large gains simply from tuning on *GD*. Tuning on *IGD* or *GD+IGD* is slighlty better. The most dramatic reduction in perplexity and WER comes from training and tuning with the near-domain data, where the inclusion of *IGD* also improves performance. With text selection, a small percentage reduces the perplexity of our LM by just under 2 points relative to including all data. We also get a small gain in word-error-rate performance relative to the best result from the models not employing selection.

P%	Use of *GD/IGD*	PPL	WER1	WER2
Baseline (no use of *GD* or *IGD*)		113.9	54.0$\star$	39.2$\star$
-	Tune *GD*	110.3	52.7	39.4
-	Tune *IGD*	109.5	52.7	39
-	Tune *GD+IGD*	109.3	53.2	38.2
-	Train+Tune CV *GD*	59.8	45.7	35.8
-	Train+Tune CV *GD+IGD*	57.7	45.3$\star$	35.5$\star$
100		55.9	44.3	36.3
50		55.9	44.2	36.9
20		54.7	44.7$\star$	35.2$\star$
10	Sel+Train+Tune CV *GD+IGD*	54.9	44.3	35.7
5		54.4	44.1	35.8
2		54.5	44.0	36.6
1		54.1	43.8$\star$	36.5$\star$

Table 4: Results of selection and LM experiments. WER1: Base WER on *DEV*. WER2: Base WER on *TEST*

5 Conclusion

For the Eveil3d project we face the challenge of creating speech recognisers for a very specific domain on unique and non-standard speech using very little data. We chose to address this problem in two ways. First, we adapted our acoustic models using MAP adaptation. Second, we expanded our near-domain language model data set through generation of invented dialogs and through intelligent selection of similar data in the out-of-domain set. While our gains over the baseline are significant, we can still see room for future work. Specifically, we would like to try bootstrapping our in-domain set for text-selection using OOD material.

References

Galliano, S.; Gravier, G.; and Chaubard, L. 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech*, volume 9, 2583–2586.

Kent, R. D. 1976. Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of Speech, Language and Hearing Research* 19(3):421.

Lawson, A. D.; Harris, D. M.; and Grieco, J. J. 2003. Effect of foreign accent on speech recognition in the nato n-4 corpus. In *Proceedings of Eurospeech*, 1505–1508.

Moore, R. C., and Lewis, W. D. 2010. Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, 220–224.

Morgan, J. J. 2004. Making a Speech Recognizer Tolerate Non-Native Speech through Gaussian Mixture Merging. In *InSTIL/ICALL 2004 Symposium on Computer Assisted Learning*.

Potamianos, A., and Narayanan, S. 2003. Robust recognition of children's speech. *Speech and audio processing, IEEE Transactions on* 11(6):603–616.

Stolcke, A. 2002. SRILM - an extensible language modeling toolkit. In *ICSLP*.

Van Doremalen, J.; Cucchiarini, C.; and Strik, H. 2010. Optimizing automatic speech recognition for low-proficient non-native speakers. *EURASIP Journal on Audio, Speech, and Music Processing* 2010:2.

Venkataraman, A., and Wang, W. 2003. Techniques for effective vocabulary selection. *Arxiv preprint cs/0306022*.