BILINGUAL AND DIALECTAL ADAPTATION AND RETRAINING

Ulla Uebler, Michael Schüßler and Heinrich Niemann

FORWISS - Bavarian Research Center for Knowledge Based Systems
D-91058 Erlangen, Germany
E-mail: {uebler,schuessler,niemann}@forwiss.de

ABSTRACT

In this paper, we report our investigations on the use of adaptation and retraining in our bilingual (Italian, German) and multi-dialectal recognition system. Our approach for bilingual speech recognition is to assume the two languages as being one, which is best suited for a task where Italian and German natives speak both languages, resulting in a variety of accents and dialects.

We performed adaptation on single speakers and speaker groups built from combinations of spoken and native language. Furthermore, we performed retraining on partitions of the adaptation or training data.

Our experiments led to an error rate reduction in all cases: compared to the baseline system, we achieved an overall improvement of 14, 12–14 and 7 % for speaker adaptation, speaker group adaptation and retraining, respectively.

Furthermore, we found among others that performance is rather stable for Italian between adaptation and retraining, while adaptation for German outperforms retraining by far.

1. Introduction

In the EU funded SPEEDATA[1] project [1], the task of data-entry in two different languages is developed. Speakers have either Italian or German as their native language and speak the other language with a certain amount of accent. Speakers come mostly from the area of South Tyrol and show a big variety of dialects especially in the German language. Further on we call the variation of a non-native speaker accent and the one of a native speaker dialect.

There are different approaches to multi- and bilingual speech recognition, for example [3, 4, 11, 12]. One approach consists of the development of a system that recognizes one language at a time, but with language independent algorithms that cover language specific aspects like homophones or coarticulation effects. A second approach is the portation of a recognizer to another language with as little retraining as possible. A third task is to recognize two or more languages at a time.

Bilingual recognition in our approach means to assume the two languages as being one. The recognizer is developed like a monolingual recognizer as far as possible, e. g. acoustic units are shared. Furthermore, words that exist in both languages with the same pronunciation like ”in” are represented only once. The lexicon of the *one* recognizer contains both Italian and German words. Some distinction between the languages is only done at the level of language modelling. The performance of the bilingual recognizer without adaptation showed to be higher than of the monolingual recognition systems [1], due to the variation in accents and dialects of the speakers of both languages.

When some speakers speak with an accent or a dialect, additional algorithms have to be applied to further improve performance, that is to adapt the system towards a speaker or a speaker group. Different experiments on adaptation with dialectal speakers have been done by [3]. In the following, there will be a short description of the baseline system and the involved adaptation and retraining algorithms. Then, experiments and results will be shown.

2. Baseline system

The employed recognition system for these experiments is adapted from a recognition system that is developed at the chair for pattern recognition at the University of Erlangen, Germany [9, 5].

During the recognition process, 12 mel cepstrum features are calcuted as well as 12 time derivatives, where the normalized energy is taken instead of the first value, resulting in 24 features for semi-continuous Hidden Markov Models. For acoustic modelling we use the technique of polyphones [9]. In this approach, different acoustic units are modelled for each phoneme depending on the context and the occurence of this context.

The data entered into the SPEEDATA system are of different types like numbers, proper names or descriptions in complete sentences.

Language models are produced according to the data fields: word lists are used for the entry of proper names, grammars for numbers and dates, statistically trained language models for whole sentences. Only these language models separate the two languages and the data fields, since one language model covers only one data field of one language.

Our data set is described in table 1. For the baseline system without adaptation, the adaptation set is used for validation during the

[1] this work is supported by the European Commission, Telematics Application Programme, project reference number LE 1999

training process. Natives in each language as well as gender are distributed equally: there are 40 speakers in the training set, 20 of them have German and Italian as their mother tongue, respectively. In each of these subsets, there are 10 female and 10 male speakers. For adaptation and test, 8 new speakers are employed with an equal distribution. The speakers of the adaptation and test set are identical. The texts of the test data are translations of the other language.

length in minutes	italian	german
training	347	399
adaptation	36	39
test	73	77

Table 1: Data set

The baseline system employs 87 bilingual phonemes. Using the technique of polyphones [8], 2000 polyphones are used for recognition. The German recognition lexicon consists of 4030 words, the Italian one of 3974, the bilingual recognition lexicon consists of 6031 words: words that occur in both languages (mostly proper names) are only modelled once.

Additionally, two monolingual recognition systems are also trained with 49 phonemes for Italian and 62 for German, leading to around 900 polyphones for each recognizer.

3. Adaptation

It has been shown that adaptation of a speaker independent system to a speaker can increase recognition performance significantly. Several adaptation methods have been proposed in the last few years. Of these, approaches which maximize the Maximum Likelihood (ML) or the Maximum a posteriori (MAP) criterion have received most attention and gave good performance [7].

We experimented with two different adaptation methods. The first one performs a common linear transformation of all codebook mean vectors $\boldsymbol{m}_k$:

$$\hat{\boldsymbol{m}}_k = \boldsymbol{A}\boldsymbol{m}_k + \boldsymbol{b}. \quad (1)$$

The transformation parameters $\Theta = \{\boldsymbol{A}, \boldsymbol{b}\}$ are estimated using the Maximum Likelihood criterion [10]:

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} \mathrm{p}(X|\Theta). \quad (2)$$

The second method is a Bayesian reestimation of the mean vectors:

$$\widehat{\boldsymbol{m}}_k = \frac{\tau_k \boldsymbol{m}_k + \sum_{t=1}^{T} \zeta_t(k)\, \boldsymbol{x}_t}{\tau_k + \sum_{t=1}^{T} \zeta_t(k)}, \quad (3)$$

where $\zeta_t(k) = \mathrm{p}(\omega_t = k|\boldsymbol{x}, \boldsymbol{\lambda})$ is computed by the forward-backward algorithm and τ_k influences adaptation speed.

In a different application, we experienced that a combination of both methods performs better than any of the methods alone [10].

4. Retraining

In cases where plenty of adaptation material is available, it seems more convenient to simply perform some training iterations on the speaker independent system using the adaptation data.

The above described adaptation algorithm changes only the codebook mean vectors, so this limited number of parameters can be estimated from few observations. The consequence is of course a limited ability to model the individual speaker. This limitation is not given for retraining, since both the HMM parameters and the codebook are reestimated. However, a larger amount of data is required in order to obtain good results.

5. Experiments

In the experiments we use two types of speaker independent baseline systems as reference for the performance evaluation: the system of two monolingual recognition systems and the bilingual recognition system.

Each of the experiments is evaluated according to the overall recognition result in both languages, the recognition of sentences in each language (li, lg), furthermore to 4 speaker subgroups, which result of the combination of the spoken language (li-italian, lg-german) and the native language of the speaker (ni-italian, ng-german), e. g. *ni-li*, see also figure 1. The performance of native speakers in both languages is also given (ni,ng).

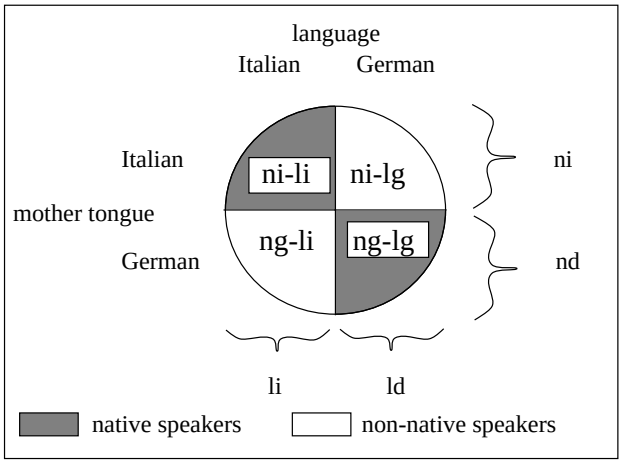

Figure 1: Speaker groups according to spoken language and mother tongue

The speech of non-native and dialectal speakers typically shows systematic differences to the standard pronunciation of German, see also Table 2. These differences also vary due to a different degree in dialect and error-prone pronunciation: inter-speaker variation (among speakers) and intra-speaker variation (in the speech of the same speaker). The influence of those effects on the system performance is also studied.

correct	spoken	example	correct	spoken	example
st	St	*Dienstag*	St	st	*Station*
h	-,x	*heute*	ts	dz	*zehn*
2	E	*möglich*	v	OU	*West*
y:	I	*über*	o:	O	*Rose*

Table 2: Pronunciation variants in German (SAMPA)

word error rates in %		total	li	lg	ni-li	ng-li	ni-lg	ng-lg	ni	ng
no adaptation	bilingual	11.27	8.92	14.34	8.55	9.28	14.97	13.71	11.34	11.20
	monolingual	12.45	8.71	17.32	7.50	9.92	18.15	16.50	12.12	12.77
adaptation (# groups)	bilingual (16)	9.70	6.67	13.65	5.55	7.79	15.12	12.19	9.70	9.69
	monolingual (8)	8.61	6.01	11.99	5.10	6.93	13.11	10.87	8.58	8.63
group adaptation (# groups)	speaker+any language (8)	9.77	6.80	13.65	5.78	7.82	14.87	12.43	9.72	9.82
	lang.+mother tongue (4)	9.93	6.93	13.85	5.93	7.94	14.58	13.12	9.68	10.18
	language (2)	9.74	6.86	13.51	5.78	7.94	14.92	12.09	9.75	9.74
	mother tongue (2)	9.89	6.95	13.73	5.93	7.97	15.07	12.38	9.89	9.89
	monol. + mother t. (2)	10.06	7.20	13.80	5.55	8.84	14.29	13.31	9.34	10.78
retraining	HMM reest.	12.49	7.55	18.94	7.58	7.53	23.68	14.19	14.56	10.42
	HMM + CB reest.	10.99	6.84	16.42	6.60	7.08	18.69	14.15	11.85	10.14
	150 min.+HMM+CB	10.45	7.12	14.80	7.09	7.15	14.68	14.93	10.38	10.52

Table 3: Word error rates for the baseline systems, adaptation and retraining experiments

5.1. Performance of the baseline systems

For the speaker independent systems, it has already been shown that the bilingual recognizer outperforms the monolingual recognizers in general [1]: the performance of the monolingual recognizer is a little higher for Italian than of the bilingual recognizer, but the bilingual recognizer performs a lot better for German than the monolingual recognizer, see the first part of table 3 (no adaptation).

Comparing the performance of the two languages, it can be seen that Italian is better recognized than German: on the one hand, Italian is a language that is easier to recognize, see also [2]. On the other hand, there is a bigger variation in German, due to different dialects and a wide range in the intensity of a dialect in comparison to standard German [6]. Furthermore, some native Italians show an immense accent when speaking German.

Within a language, of course, natives are recognized better than non-natives who have a larger (inter- and intra-speaker) variety than natives.

5.2. Adaptation

All experiments are carried out with the combined adaptation algorithm. Its parameters were determined in preparatory experiments.

The first experiment in this series is performed with the bilingual recognizer with an adaptation per speaker and language which leads to an adaptation of 16 groups (8 speakers x 2 languages). Furthermore, adaptation is also carried out with the two monolingual recognizers. Results are shown in the second part of table 3 (adaptation). In this experiment, the error rate decreased by 14 %.

Looking at the performance of the monolingual recognizers, a big improvement can be seen with a 31 % smaller error rate. Now, the monolingual recognizers even perform better than the bilingual recognizer. It is surprising that the adaptation algorithm is more effective for the monolingual than for the bilingual recognizer, resulting in a better overall performance.

The next experiment deals with building groups for adaptation: if adaptation is not performed by speaker and language but in larger groups, there are more adaptation data available for one adaptation step, but the adaptation data is less specific for the group being adapted.

From the initial 16 groups (8 speakers, 2 languages) new groups are built according to speaker or language characteristics, i. e. speaker, spoken language, or mother tongue of the speaker, leading to 8, 4 or 2 groups per adaptation step. Results are shown in the third part of table 3 (group adaptation).

Recognition rates do not vary very much, although it can be seen that best results are obtained when adapting to a speaker (any language) or adapting to a language. Within these experiments on group adaptation, it is astonishing that the performance with four groups is slightly worse than with only two groups, since the characteristics of the language should be represented best.

With regard to the languages, in most cases Italian improves more than German: with the bilingual recognizer performance for the German language is improved by 5 %, for Italian by 25 %.

When adapting to both speaker and language, performance decreases for the group *ni-lg*, which is the group with the auditively biggest inter- and intra speaker variation. In this particular case, the adaptation algorithm does not achieve an improvement, possibly due to the high variation caused by accent and dialect, i. e. intra-speaker variation. This effect can be compensated by adapting to larger groups. The biggest improvement with respect to the baseline system for this group is found with an adaptation to language and mother tongue speakers.

With the monolingual recognition systems, there is only one system of 2 subgroups to be built – adapting to the mother tongue of the speaker (monolingual + mother tongue in table 3). The performance achieved with this experiment is worse than adapting the monolingual recognizers to each speaker (10.06 %, per speaker: 8.61%). It is also slightly worse than the performance of the bilingual recognizer for the subgroups, although best performance within those subgroups is achieved for native Italians (*ni* and native Italians speaking German *ni-lg*.

The group that obtained the highest improvement during adaptation is native Italians speaking Italian: this is also the group with the least inter- and intra-speaker variation due to dialect and accent.

5.3. Retraining

The first set of experiments is carried out with retraining on the adaptation data, i. e. on a partition of 75 minutes of speech depending on the subgroup that is retrained, see table 3 (retraining).

In a first step, retraining with the adaptation data is performed by reestimation of the HMM parameters, with and without codebook reestimation. Without reestimation of the codebook the performance is worse than the baseline system for all introduced subgroups.

Performing a reestimation of the codebook and another retraining of the HMM parameters only improves performance slightly when retraining on the four subgroups (*ni-li* etc.). This is still worse than any result of the adaptation experiments. This result is caused by the bad performance of German. Italian, however, performs in the same range as with adaptation.

For these four groups, retraining is performed also on a partition of the training set, which matches the group characteristics, e. g. native Italians speaking German with around 150 minutes training material per group. Best results are achieved with codebook and HMM parameter reestimation. These results are 7 % better than those of the baseline systems, but still worse than with adaptation. When evaluating the performance per speaker group it can be seen that, in relation to the adaptation experiments, native speakers (*ni-li, ng-lg*) are recognized worse whereas for the non-native groups (*ng-li, ni-lg*) the bilingual and adapted recognizer is outperformed by 3 to 8 %. The monolingual adapted recognizer is still the best for all speaker groups.

6. Conclusion

In our experiments in the domain of bilingual speech recognition, adaptation decreases the error rate for our bilingual recognizer by 14 %. For the monolingual recognizers the error rate was reduced by 31 % resulting in an overall error rate of 8.61 %.

Building speaker groups according to characteristics like language, mother tongue and the combination of both, achieves nearly the same performance like adaptation to a single speaker.

We also retrained the system with those speaker groups. Evaluation on the groups gave an improvement as well. When retraining is performed on that partition of the training set which matches the group characteristics, although there are no common speakers to both sets, the error rate is reduced by 7 % with respect to the baseline system.

7. REFERENCES

1. U. Ackermann, F. Brugnara, M. Federico, and H. Niemann. Application of Speech Technology in the Multilingual SpeeData project. In *3rd Crim-Forwiss Workshop*, Montréal, 1996.

2. J. Barnett, A. Corrada, G. Gao, L. Gillick, Y. Ito, S. Lowe, L. Manganaro, and B. Peskin. Multilingual Speech Recognition at Dragon Systems. In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, USA, 1996.

3. J. Billa, K. Ma, J. McDonough, G. Zavagliakos, D. Miller, K. Ross, and A. El-Jarouodi. Multilingual Speech Recognition: The 1996 Byblos Callhome System. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 363–366, Greece, September 1997.

4. P. Bonaventura, F. Gallochio, and G. Micca. Multilingual Speech Recognition for Flexible Vocabularies. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 355–358, Greece, September 1997.

5. T. Kuhn. *Die Erkennungsphase in einem Dialogsystem*, volume 80 of *Dissertationen zur künstlichen Intelligenz*. infix, St. Augustin, 1995.

6. H. Moser. Methodische Überlegungen zur Untersuchung des gesprochenen Deutsch in Südtirol. In H. Moser, editor, *Zur Situation des Deutschen in Südtirol*. Innsbrucker Beiträge zur Kulturwissenschaft, Innsbruck, Österreich, 1982.

7. L. Neumeyer, A. Sankar, and V. Digalakis. A comparative study of speaker adaptation techniques. In *Proc. European Conf. on Speech Technology*, pages 1127–1130, Madrid, 1995.

8. E. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Acoustic Modelling of Subword Units in the ISADORA Speech Recognizer. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 577–580, San Francisco, 1992.

9. E. G. Schukat-Talamazzini. Automatische Spracherkennung. Habitilation, Technische Fakultät der Universität Erlangen-Nürnberg, Erlangen, 1993.

10. M. Schüßler, F. Gallwitz, and S. Harbeck. A Fast Algorithm for Unsupervised Incremental Speaker Adaptation. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 1019–1022, München, 1997.

11. F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke. A Study of Multilingual Speech Recognition. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 359–362, Greece, September 1997.

12. S. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.-L. Gauvain, D. Kershaw, L. Lamel, D. Leeuwen, D. Pye, A. Robinson, H. Steeneken, and P. Woodland. Multilingual large vocabulary speech recognition: the European SQUALE project. *Computer Speech & Language*, 11:73–89, 1997.