Cross-lingual Portability of Chinese and English Neural Network Features for French and German LVCSR

Christian Plahl, Ralf Schlüter and Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Department
RWTH Aachen University, Aachen, Germany
{plahl,schlueter,ney}@cs.rwth-aachen.de

***Abstract*—This paper investigates neural network (NN) based cross-lingual probabilistic features. Earlier work reports that intra-lingual features consistently outperform the corresponding cross-lingual features. We show that this may not generalize. Depending on the complexity of the NN features, cross-lingual features reduce the resources used for training —the NN has to be trained on one language only— without any loss in performance w.r.t. word error rate (WER). To further investigate this inconsistency concerning intra- vs. cross-lingual neural network features, we analyze the performance of these features w.r.t. the degree of kinship between training and testing language, and the amount of training data used.**

Whenever the same amount of data is used for NN training, a close relationship between training and testing language is required to achieve similar results. By increasing the training data the relationship becomes less, as well as changing the topology of the NN to the bottle neck structure. Moreover, cross-lingual features trained on English or Chinese improve the best intra-lingual system for German up to 2% relative in WER and up to 3% relative for French and achieve the same improvement as for discriminative training. Moreover, we gain again up to 8% relative in WER by combining intra- and cross-lingual systems.

***Index Terms*—neural network, cross-lingual portability, feature extraction, LVCSR**

I. Introduction

Several state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems based on subsystems with different acoustic front-ends are developed at RWTH [1], [2]. In order to benefit from the different acoustic front ends, a complete GMM/HMM based training has been performed independently for each front-end and language. Since neural network (NN) probabilistic features have become a major component of state-of-the-art speech recognition systems, NNs have been trained independently for each language, too. In the last years several investigations to further improve the probabilistic NN features have been performed, resulting in a very complex structure and high computational costs for training these NNs [3], [4], [5]. Whereas the training of NNs is extremely time consuming, NN decoding is very efficient. Therefore, decoding a previously trained NN to provide acoustic probabilistic features for GMM/HMM training and for the recognition step is an efficient reuse of resources available. By this, we reduce the computational costs and resources in the NN feature extraction process, without any loss in performance w.r.t. word error rate (WER), whereas in [6] the acoustic training has been simplified by combining different feature streams by neural networks.

In [7] Tandem-PLP NN posterior features, trained on 1800 hours of English data, have improved the recognition results for Arabic and Chinese. Nevertheless, the best results in [7] have been obtained when the NN features are trained on Chinese or Arabic, even when a small amount of less than 100 hours are used to train the NN and the acoustic model. By adapting the weights of a previously trained NN to another language the NN benefits from a good initialization, e.g. when only a small amount of training data is available. Moreover, a good initialization of the NN could save necessary resources during training [8]. In [9] a developed Hungarian system is improved by adapting English trained Tandem-PLP NN features to Hungarian, where 2000h of English data have been provided for NN training. Again, a very small amount of intra-lingual data (7h) has been available only. Overall, a large amount of data of another language as for the acoustic training is used for NN training, whereas the acoustic model is trained on a small amount of data only.

In contrast, in this paper we concentrate on the questions of this inconsistency of NN training and testing language concerning:

- the role of the structure/topology of the neural network —only Tandem-PLP probabilistic features have been used for acoustic training so far,
- the degree of kinship of training and testing language for the (cross-lingual) NN probabilistic feature extraction,
- and the dependency on the amount of data used for training the NNs.

Therefore, we have trained several NNs with hierarchical structure and with and without bottle neck topology, on Chinese, English, French and German [5], [3]. The corresponding NN probabilistic features are evaluated on the Quaero task

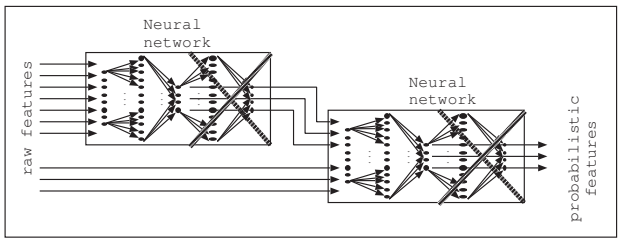

Fig. 1. Hierarchical processing of neural networks. Here the neural networks consist of the bottle neck structure, where the output of an inner layer is taken during decoding instead of the posteriors from the output layer. In training, all five layer are used.

for French and German. We show that the topology of the NN chosen is more significant than the specific training language, since almost all NN features achieve similar results, irrespective of whether training and testing language of the NN matches. Moreover, we obtain the best results for French and German when we use the (cross-lingual) NN features obtained by forwarding the data to a NN trained on Chinese or English. In addition, the time for training NN features for each language is reduced significantly, because the time and resources consuming NN training have to be performed on one language only, here, to obtain best results, Chinese. After training, the French and German training and testing data are forwarded to the final NN and cross-lingual NN features are provided. These cross-lingual features could be used for GMM/HMM training without any loss in WER.

Finally we show, whenever different intra- and cross-lingual NN features without any extra costs are available, the systems trained on these features produce different complementary errors. Moreover, we gain slightly more by the combination of the different systems as shown in other system combination experiments on the same corpus.

II. Neural Network Feature Extraction

In this section, we briefly summarize the different NN topologies used in the probabilistic NN features extraction step. The NN feature extraction step is based on the hierarchical processing of NNs and on the bottle neck structure. The hierarchical bottle neck topology used is the result of an exhausting investigation in the field of NN features in the last years [4], [5], [3]. In [10] a large comparative study of several NN features like Tandem-PLPs, features based on TempoRAl Pattern (TRAPs), their further development to hidden activation TRAPs (HATs) and NN features based on multi-resolution RASTA filtering have been performed. In [3] the best approach is further developed resulting in the very complex hierarchical bottle neck topology, Figure 1. The training of such a hierarchical NN requires high computational resources and therefore reducing this cost is of high interest. Next, we will shortly summarize the input features used and the hierarchical bottle neck topology.

A. Multi-resolution RASTA Features

In [11] an extension of the RASTA, the multi-resolution RASTA (MRASTA) filtering is introduced. The filters are realized by two dimensional band-pass filters using separate ranges of modulation frequencies to extract a set of multiple resolution filters. The RASTA filtering itself has been proposed as a modification of the TRAP-based probabilistic features, introduced in [12]. We use the MRASTA features as input to train each of our NNs.

In order to extract the MRASTA features, 19 critical bands of the auditory spectrum, extracted from the Fourier transform of a signal every 10 ms are taken. Next, each critical band is filtered with a bank of several low-pass filters represented by six first derivatives and six second derivatives of Gaussian functions.

B. Hierarchical Bottle Neck Processing

The concept of hierarchical bottle neck features combines the advantages of the hierarchical processing of NNs and the bottle neck topology [3]. As shown in Figure 1 the hierarchical processing consists of a cascade of NNs, where the NNs use the decoded features of a previous trained NN as input [3], [13]. Such a hierarchical processing improves the accuracy of the final posterior estimates and the complete ASR system trained on the improved probabilistic features. On the other hand, the goal of bottle neck features is to provide the ability to compress the input raw features in an arbitrary size and to ensure a good class separability of the output features [5].

In the hierarchical bottle neck processing the 3-layer NN of the hierarchical processing is exchanged by the bottle neck NN [3]. In order to provide the necessary modeling power the first hidden layer uses 4000 nodes, whereas the bottle neck, second hidden layer, is set equal to the number of output units. This allows a direct comparison between the final posterior features and the probabilistic features obtained from the bottle neck. In order to further improve the classification error rate, the last hidden layer has been enlarged again to 2000 nodes.

After training the first NN with the fast modulation frequencies of the MRASTA features, the second NN (NN2) takes the linear output of the bottle neck and the slow modulation frequencies of the MRASTA processing as input. Next, the linear output of the bottle neck of NN2 is normalized by mean and variance normalization. Finally, the features are transformed by PCA and are reduced to a dimensionality of 30. In the experimental section these features are referred to as *Hier.bn.mrasta*. *Hier.mrasta* features are extracted by the hierarchical processing without the bottle neck topology.

III. Cross-lingual Neural Network Features

In the last years, the topology of the neural network, as well as the input features, have been under investigation, resulting in a very complex, resources and time consuming NN training step. Nevertheless, the decoding of such a NN is very efficient. Therefore, instead of training NNs for each database or language, our purpose is to train NN probabilistic features on one database or language only and to reuse the

trained NN in the decoding step to produce NN probabilistic features for other databases as well. When the training and testing language differ, the resulting NN features are referred as cross-lingual NN features. Intra-lingual NN features are extracted, when training and testing language are identical.

As shown in [7], [9], porting a previously trained NN to other languages helps to improve the system performance. There, the best results are obtained when intra-lingual NN features are used, even though the cross-lingual features are trained on a huge amount of data. In contrast to the experiments done in [7], we have exchanged the Tandem-PLP features by the complex Hier.mrasta and Hier.bn.mrasta features, Section II-B. We have chosen a similar amount of data to train the different NNs to analyze the relationship between training and testing language. In our case, porting NN features save computational costs and resources. Moreover, we could simplify the development process of several speech recognition systems for multiple languages by training a NN for one language only. By decoding, the NN provides the cross-lingual NN features.

In order to investigate the effect of the different NN features, we have extracted intra-lingual NN features and cross-lingual NN features for two European languages. After NN training, we generate the cross-lingual NN features for French and German by forwarding the data to the NNs trained on Chinese or English, and the intra-lingual features by the NNs trained on French and German. Again, we have taken the two languages Chinese and English for training the NNs and to decode the cross-lingual NNs to analyze the influence of the relationship of training and testing language.

While the relationship for Chinese to French and German is not as large as for English, we create, for comparison, the Hier.mrasta and Hier.bn.mrasta NN features on a medium and a large corpus. As shown in the experimental section, even the medium corpus achieves good results. In Chinese tonal information play an important role to distinguish tonemes, phonemes with tonal information, and words [14]. Therefore, all Mandarin ASR systems cope with these information [2]. In European languages this information is not helpful to improve the system performance. The Chinese system used is described in detail in [2]. For comparison, the final NN features are reduced by PCA to a 30 dimensional feature vector.

As in [3], the English Hier.mrasta and Hier.bn.mrasta NN features are trained on 310h of speech data. The 44 phonetic targets are derived from a forced alignment of a previously trained English system. The final Hier.mrasta and Hier.bn.mrasta NN features are reduced to 95% of the variability by PCA to a 30 dimensional feature vector.

IV. Acoustic Modeling

In order to evaluate the different NN features, several ASR systems based on GMM/HMM models are trained. The systems built are based on the French and German systems used in the QUAERO 2010 Evaluation [1], competitive to current systems within the project. The acoustic models are trained with the RWTH speech recognition system [15].

A. Acoustic Features

Compared to [1], the acoustic front-ends of the systems for French and German differ in the type of the NN probabilistic features only. On one hand, the training is based on two different types of NN topologies, the hierarchical processing and the bottle neck structure, see Section II. On the other hand, the training of the NN is performed on different training data and languages. When training and testing languages of the NN match, intra-lingual NN features are produced, cross-lingual NN features, when a mismatch between the training and testing languages exist. The final models are trained on French and German. The extracted NN features are augmented with classical short-term MFCC features resulting in a final feature stream of 75 components. The MFCC features have been normalized by segment-wise mean and variance normalization. Furthermore, before augmented with the NN features, all MFCC features within a sliding window of length nine are concatenated and projected to a 45 dimensional feature space using a linear discriminant analysis (LDA).

B. Acoustic Training

The acoustic models for all systems are based on triphones with cross-word context, modeled by a 3-state left-to-right hidden Markov model (HMM). A decision tree based state tying is applied resulting in a total of 4500 generalized triphone states. The acoustic models consist of Gaussian mixture distributions with a globally pooled diagonal covariance matrix.

In order to compensate for speaker variations constrained maximum likelihood linear regression speaker adaptive training (SAT/CMLLR) is performed. In addition, during recognition, MLLR is applied to the means of the acoustic models. For computational reason we disclaim a full training including discriminative training.

V. Corpora

The NN probabilistic features have been trained on Chinese, English, French and German. In contrast, the training of the acoustic models is performed on the French and German data only. All data consist of broadcast news (BN) and broadcast conversation (BC).

The training data of the two corpora for Chinese have been collected by LDC and consist of audio data of the first four years of the GALE project (releases P1R1-4, P2R1-2, P3R1-2, P4R1). Whereas the large corpus is built from 1600h of speech data from all releases, the medium corpus consists of 230h from the P1R1-4 data only.

Approximately 310 hours of American BN of speech data are used for training the English probabilistic NN features. The whole corpus consists of 140 hours of HUB4 speech data and 170 hours of TDT4 data collected by LDC.

The training corpus for French consists of 140h of Ester1 and Ester2. In addition, 90h of speech data collected from various sources within the Quaero project are used for training. The French NN probabilistic features as well as the GMM/HMM training for French has been performed on the full corpus of 230 hours.

For German 50h of acoustic data are used. As for French, this data are provided within the Quaero project and consists of BC data collected from various sources of the web.

A more detailed description of the training data for Chinese is given in [2], in [3] for English and for French and German in [1].

TABLE I
ACOUSTIC DATA OF THE QUAERO DEVELOPMENT AND EVALUATION CORPORA FOR FRENCH (FR) AND GERMAN (DE). THE DEVELOPMENT CORPUS OF 2010, MARKED BY *, IS USED FOR TUNING THE PARAMETERS OF THE FINAL ASR SYSTEMS.

		Tuning and testing data		
		dev10*	eval10	eval09
FR	total data	3.8h	2.7h	3.9h
	# segments	2478	1866	1067
DE	total data	3.7h	3.4h	3.7h
	# segments	1207	1126	1356

The evaluation of the systems is performed on the Quaero development and evaluation data of 2009 and 2010 for French and German. While the systems are tuned on the development corpus, marked by * in Table I, the evaluation corpora are used for testing only. Within the Quaero project, a transcription and a manual segmentation of these corpora are provided. As for the Quaero training data, the development and evaluation corpora consist of a mix of different speech sources, containing broadcast news, broadcast conversation, several pod cast shows and other speech data collected from the web. Table I summarizes the development and testing corpora used.

In order to cope with high lexical variety in the German language, a large number of distinct lexical forms can be generated by derivation, compounding, and inflection, words have been decomposed into sublexical fragments [1], [16]. After decoding, a preprocessing step joins the sublexical fragments again.

VI. EXPERIMENTS

We have trained several recognition systems for French and German using probabilistic features extracted by NNs trained on Chinese, English, French and German to analyze the effect of cross-lingual portability of the NN probabilistic features. In addition, we compare the cross-lingual NN features and the intra-lingual NN features. As reported in the literature, applying any NN features helps to improve the system performance in terms of WER for all languages and corpora. We use a two pass decoding system, where we apply the speaker adapted model in the second pass.

A. French

As shown in Table II all NN features could gain over the baseline system, and all systems using Hier.bn.mrasta features outperform the corresponding systems based on Hier.mrasta features. Moreover, the systems S2, S3 and S4 achieve similar results, even so the training and decoding language of the NN does not match. All results are within a range of 1% relative and the amount of data used to train the NNs are approximately the same. Here, the bottle neck structure produces language independent features and provide a good global structure of speech production, tied over different languages. The English cross-lingual feature could benefit from its close relationship of the European languages to each other, and achieve always slightly better results as the corresponding Chinese cross-lingual trained on a similar amount of data, here 230h.

TABLE II
RECOGNITION RESULTS FOR FRENCH AFTER SPEAKER ADAPTATION USING CROSS-LINGUAL AND INTRA-LINGUAL NN FEATURES. THE NN TRAINED ON CHINESE (CN), ENGLISH (EN) AND FRENCH (FR) ARE USED TO PRODUCE THE NN FEATURES FOR FRENCH TO TRAIN A GMM/HMM ASR SYSTEM. THE TOPOLOGIES USED TO TRAIN THE NN ARE THE HIERARCHICAL MRASTA (HIER.MRASTA) AND THE HIERARCHICAL BOTTLE NECK MRASTA (HIER.BN.MRASTA) PROCESSING.

NN-train language	NN feature type		WER[%] dev10*	eval10	eval09
FR	(no NN features)		24.1	25.4	34.2
	Hier.bn.mrasta	(S3)	23.1	23.7	33.4
EN	Hier. .mrasta		23.5	24.0	33.2
	Hier.bn.mrasta	(S2)	23.0	23.9	33.2
CN	Hier. .mrasta.230		23.7	24.3	33.6
	Hier.bn.mrasta.230	(S4)	23.3	24.1	33.3
	Hier. .mrasta.1600		23.1	24.1	33.1
	Hier.bn.mrasta.1600	(S1)	**22.4**	**23.5**	**32.7**

The best recognition result could be achieved by the Hier.bn.mrasta features trained on 1600 hours of Chinese data (S1). The huge amount of data could reduce the effect of the degree of kinship of the training and decoding language. This could also be verified for the structure of the neural network. The Hier.mrasta features trained on 1600 hours of Chinese data perform as good as the systems S2, S3 and S4. Therefore, the combination of huge amount of data combined with the topology of the neural network results in an absolute improvement of 0.7% in WER by the Chinese cross-lingual features compared to the intra-lingual features on dev10 and eval09. Moreover, the improvement of 0.7% absolute in WER is reported as improvements of discriminative training in Table 5 in [1]. Therefore, we expect a further improvement of the system when discriminative training is applied.

These results show that cross-lingual features are not always worse than intra-lingual features. In this case, cross-lingual features could even outperform the intra-lingual features. The main reason for this is the large amount of training data used and the bottle neck structure of the NN. Increasing the amount of training data results in a more robust estimation of the weights of the NN. Since the same amount of data is sufficient to achieve similar results for cross- and intra-lingual features when the relationship of training and testing languages is close, the bottle neck structure gains most. System S1, the NN is trained on Chinese, is improved by more than 0.6% absolute by the bottle neck structure for all corpora. The bottle neck structure is not only relevant for a good class separability and to provide a good and compact representation of the input features, the bottle neck structure focuses on speech production aspect, common across different languages. This is supported by the fact, that only the cross-lingual bottle neck features could gain over the intra-lingual features. Cross-domain and cross-system adaptation effects play an insignificant role only.

B. German

Although, the amount of training data for the training of the acoustic GMM/HMM model for German is smaller than for the French task, we get similar results. Results are shown in Table III. Again, system S3 and S4 and the system based on Chinese Hier.mrasta features, trained on 1600 hours, perform equally well on the dev10 and eval09 corpora, whereas on eval10 a gap of 0.5% exists. This is due to the effect of the degree of similarity of German and Chinese.

TABLE III
CROSS-LINGUAL AND INTRA-LINGUAL RECOGNITION RESULTS FOR GERMAN AFTER SPEAKER ADAPTATION. THE NNS TRAINED ON CHINESE (CN), ENGLISH (EN) AND GERMAN (DE) ARE USED TO DECODE NN FEATURES FOR GERMAN TO TRAIN THE FINAL GMM/HMM ASR SYSTEMS. THE TOPOLOGIES USED TO TRAIN THE NN ARE THE HIERARCHICAL MRASTA (HIER.MRASTA) AND THE HIERARCHICAL BOTTLE NECK MRASTA (HIER.BN.MRASTA) STRUCTURE.

NN-train language	NN feature type		WER[%] dev10*	eval10	eval09
DE	(no NN features)		18.6	18.1	27.1
	Hier.bn.mrasta	(S3)	17.8	17.2	26.2
EN	Hier. .mrasta		17.9	17.3	26.2
	Hier.bn.mrasta	(S2)	**17.4**	**17.0**	**25.4**
CN	Hier. .mrasta.230		18.3	18.1	26.8
	Hier.bn.mrasta.230	(S4)	17.9	17.9	26.4
	Hier. .mrasta.1600		18.0	17.7	26.3
	Hier.bn.mrasta.1600	(S1)	**17.4**	**17.0**	**25.5**

In contrast to the experiments on French, S2 performs as good as S1. The system S2 performs so well on the German task because the training domain of the German training and testing data and the English data mostly consist of broadcast news, whereas for the French task other domains are included as well, e.g. podcasts from various sources of the web.

The best recognition performance of 17.4% on dev10 is achieved by the Hier.bn.mrasta features trained on 1600 hour of Chinese speech data (S1). This is a relative improvement of over 2% in WER of the intra-lingual system (S3) on dev10 and 6.5% compared to the baseline system, which uses no NN features at all. As shown in Table III, again the intra-lingual system has been beaten by the cross-lingual features. In addition, we could draw the same conclusion on the German task as done on the French task. The degree of similarity between training and testing language could be reduced by increasing the data used for NN training and by the bottle neck structure. Moreover, we could verify, that, even though each language has its own phoneme set, languages share phonetic distinction at the level of articulatory features, such as voicing, frication and nasality.

Overall, the training process for French and German can now be simplified. Instead of training two NNs for each language, the training of the Chinese Hier.bn.mrasta on 1600h is sufficient. The final cross-lingual features for German and French are extracted by forwarding the data to the previously trained Chinese NN.

VII. SYSTEM COMBINATION

In order to get a better analysis on how complementary the intra-lingual and cross-lingual systems are, we have performed system combination based on confusion networks as described in [17] for systems S1-S4. See Table II and Table III for details about the NN training language. The lattices are converted into confusion networks and the weights for the different system are optimized on the development set. We have performed combination of two, three and four systems and the results are shown in Table IV. Other combinations like ROVER have been tested as well, resulting in slightly worse results.

The best recognition for French is achieved when we combine all four systems, whereas the improvement is small compared to the combination S1 $\oplus$ S2 $\oplus$ S3. Here, the weight for S4 is set to 0.05 for German and 0.20 for French, whereas the other weights are set equally. Whenever S1 and S4 are combined, the weight for S4 is less. This is not the case when S4 is combined with the other two systems. Since S2, S3 and S4 perform equally well for French we get similar results for the two and three system combinations and equal combination weights for each system. Moreover, we get the same result also for S1 $\oplus$ S4, whereas all other combinations including S1 achieving a much lower WER. This is, because both, S1 and S4 use cross-lingual features trained on Chinese and therefore produce similar errors. This behavior has been also observed on the German task, where e.g. S1 $\oplus$ S4 gets the worst combination result of two systems.

Overall, we achieve an improvement of up to 5.5% relative on the French task, and get a final WER on dev10 which is slightly worse than the combination result in [1], but giving a better generalization on eval10. We note, that the systems here are trained without discriminative training which will result in an additional improvement of 3%-5% relative. Due to an improvement of the postprocessing step of the sublexical fragments for German, we end up with a better baseline system and a better final combination result.

VIII. SUMMARY AND CONCLUSIONS

In this paper, we ported cross-lingual NN features to French and German. We showed that such features trained on Chinese or English improved the recognition performance. Moreover, we achieved similar results by the cross-lingual NN features as for the intra-lingual features, when the NN is trained on the same amount of data. The fact that cross-lingual features were always outperformed may not generalize. Depending on the amount of training data available for the specific language and the NN topology used, the degree of kinship of training and testing language got unimportant. Even more, the good generalization of the bottle neck structure itself and the compact representation of the input features by the bottle neck had provided cross-lingual specific features. This helped to reduce the degree of kinship of the training and testing language for the cross-lingual features and to outperform the intra-lingual features.

The Chinese cross-lingual Hier.bn.mrasta NN features achieved the best recognition results on French and German

TABLE IV
SYSTEM COMBINATION RESULTS FOR GERMAN AND FRENCH. THE SYSTEMS S1, S2, S3 AND S4, EACH USING THE HIER.BN.MRASTA FEATURES, ARE COMBINED BY FRAME WISE LATTICE BASED SYSTEM COMBINATION. THE NN USED FOR FEATURE EXTRACTION ARE TRAINED ON CHINESE, ENGLISH, FRENCH AND GERMAN. THE SYSTEMS COMBINED ARE MARKED BY X. WE HAVE TESTED THE COMBINATION OF TWO, THREE AND FOUR SYSTEMS.

Systems (NN training language)				German (WER[%])			French (WER[%])		
S1 (CN)	S2 (EN)	S3 (FR/DE)	S4 (CN-230h)	dev10*	eval10	eval09	dev10*	eval10	eval09
X				**17.4**	**17.0**	**25.5**	**22.4**	**23.5**	**32.7**
	X			**17.4**	**17.0**	**25.4**	23.0	23.9	33.2
		X		17.8	17.2	26.2	23.1	23.7	33.4
			X	17.9	18.1	26.8	23.1	24.1	33.1
X	X			**16.3**	**16.0**	**24.5**	**21.5**	**22.6**	**31.6**
X		X		16.5	16.0	24.7	21.4	22.4	31.8
X			X	16.9	16.5	25.0	21.8	22.9	31.8
	X	X		16.4	16.0	24.6	21.9	22.6	32.3
	X		X	16.5	16.4	24.8	22.0	23.0	32.1
		X	X	16.8	16.5	25.0	21.9	22.6	32.2
X	X	X		**16.1**	**15.7**	**24.2**	**21.3**	**22.1**	**31.7**
X	X		X	16.2	15.9	24.3	21.5	22.4	31.8
X		X	X	16.5	16.0	24.6	21.4	22.3	31.6
	X	X	X	16.3	15.9	24.4	21.5	22.4	32.1
X	X	X	X	**16.0**	**15.7**	**24.1**	**21.3**	**22.1**	**31.5**

and improved the best intra-lingual French system up to 3% relative. Moreover, discriminative training, as reported in [1], achieved the same improvement. So we expect an additional improvement, when discriminative training will be applied. On the German task, we ended up with a relative improvement of up to 2%.

Overall, the system development circle for German and French could be simplified now, without any loss of performance w.r.t. WER. Instead of training NN probabilistic features for each corresponding language within a project, a training of Hier.bn.mrasta NN features for one language will be sufficient —here Chinese. Since the NN feature extraction was very efficient for decoding, cross-lingual NN features reduced the necessary amount of training resources used and optimized the overall training and decoding process and the resources available.

Furthermore, if different ASR systems were available, we showed that different cross- and intra-lingual NN features were complementary to each other and that the combination of the systems resulted in a further improvement of up to 6%-8% relative. Overall, we got better recognition results for both languages, as reported in [1].

In order to further improve the systems and complete the analysis of the training of cross-lingual specific features by the bottle neck structure, a detailed analysis of the error will be necessary. Moreover, to verify the current results, experiments on other European languages have to be performed as well. In addition, the resources saved will be used to improve the NN features further by increase the complexity of the NN.

ACKNOWLEDGMENTS

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

REFERENCES

[1] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A. El-Desoky Mousa, S. Hahn, D. Nolden, R. Schlüter, and H. Ney, "The RWTH 2010 quaero ASR evaluation system for English, French, and German," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 2212–2215.

[2] C. Plahl, B. Hoffmeister, G. Heigold, J. Lööf, R. Schlüter, and H. Ney, "Development of the GALE 2008 Mandarin LVCSR system," in *Interspeech*, Brighton, U.K., Sep. 2009, pp. 2107–2110.

[3] C. Plahl, R. Schlüter, and H. Ney, "Hierarchical bottle neck features for LVCSR," in *Interspeech*, Makuhari, Japan, Sep. 2010, pp. 1197–1200.

[4] F. Valente, M. Magimai.-Doss, C. Plahl, and S. Ravuri, "Hierarchical processing of the modulation spectrum for GALE Mandarin LVCSR system," in *Interspeech*, Brighton, U.K., Sep. 2009, pp. 2963–2966.

[5] F. Grézl, M. Karafiat, S. Kontar, and J. Cernock, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, Honolulu, HI, USA, Apr. 2007, pp. 757–760.

[6] C. Plahl, R. Schlüter, and H. Ney, "Improved acoustic feature combination for LVCSR by neural networks," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1237–1240.

[7] A. Stolcke, F. Grézl, M. Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptron," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Toulouse, France, May 2006, pp. 321–324.

[8] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, pp. 1771–1800, August 2002.

[9] L. Tóth, J. Frankel, G. Gasztolya, and S. King, "Cross-lingual portability of MLP-Based tandem features – a case study for English and Hungarian," in *Interspeech*, Australia, Aug. 2008, pp. 2695–2698.

[10] F. Valente, M. Magimai-Doss, C. Plahl, S. Ravuri, and W. Wang, "A comparative large scale study of MLP features for Mandarin ASR," in *Interspeech*, Makuhari, Japan, Sep. 2010, pp. 2630–2633.

[11] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 361–364.

[12] H. Hermansky and S. Sharma, "TRAPs - classifiers of temporal patterns," in *Proc. Int. Conf. on Spoken Language Processing*, Sydney, Australia, Dec. 1998, pp. 1003–1006.

[13] F. Valente, J. Vepa, C. Plahl, C. Gollan, H. Hermansky, and R. Schlüter, "Hierarchical neural networks feature extraction for LVCSR system," in *Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 42–45.

[14] X. Lei et al., "Improved tone modeling for Mandarin broadcast news speech recognition," in *Interspeech*, Pittsburgh, Pennsylvania, USA, Sep. 2006, pp. 1237–1240.

[15] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, "The RWTH Aachen University open source speech recognition system," in *Interspeech*, Brighton, U.K., Sep. 2009, pp. 2111–2114.

[16] A. El-Desoky Mousa, M. A. Basha Shaik, R. Schlüter, and H. Ney, "Sub-lexical language models for German LVCSR," *IEEE workshop on spoken language technology*, pp. 159–164, Dec. 2010.

[17] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *NIST Speech Transcription Workshop*, College Park, MD, 2000.