Internal Use Only

002

TR-IT-0261

Adaptation of BRNN Speech Recognition Systems

Petra Philips Mike Schuster

1998 年 4 月 21 日

Because speaker-independent large-vocabulary systems need huge amounts of training data the parameters of the acoustical units have a high variance and thus give poor models for individual utterances, being sensitive to changes of environment (speaker or channel). One attempt to solve this problem is to transforme the feature and/or model space in order to reduce the mismatch between the acoustical data and the acoustical models of the system.
We present some experimental results achieved with supervised and unsupervised adaptation of a hybrid BRNN (Bidirectional Recurrent Neural Network) phoneme recognition system on TIMIT data using

1. a Linear Input Network (LIN)
2. retraining the BRNN with weight-sharing

We show also how unsupervised adaptation can be improved using only a simple acoustical confidence measure based on the posterior probability of the recognized class for every frame.

ⓒＡＴＲ音声翻訳通信研究所

ⓒATR Interpreting Telecommunications Research Laboratories

目次

1 Adaptation

1.1 Motivation

State of the art speech recognizers, i.e. large-vocabulary speaker-independent systems, have to deal with huge amounts of data, which implies to deal with large variations between speakers and environment. This leads to a higher variance of the parameters of the acoustical units and to poor models for individual utterances. As a consequence, for example speaker-independent systems have higher error rates than speaker-dependent systems, even if the training data is sufficient ([17]). Also, almost all recognizers are still sensitive to changes of the environment characteristics for the training and testing data.

One attempt to solve this problem is to transform the feature and/or model space in order to reduce the mismatch between the acoustical data and the acoustical models of the system.

1.2 Terminology

One possible method for reducing the sensitivity of the system to large variations of speech characteristics is adaptation. **Adaptation** is here defined as the process of adjusting the parameters of the system (either in the feature space or in the model space) to improve the model for a new acoustic environment (speaker for example).

Examples are the transformations of the means and variances of the Gaussian mixtures of HMMs, affine transformations of the feature space for a new speaker, choosing between different trained systems, and retraining of neural networks.

In the adaptation literature the term **normalization** is often used for a special case of adaptation of the input feature space. Normalization is the process of transforming the input speech, so that it appears to the system to have the same characteristics for every acoustic environment. Speaker or environment are mapped to some trained prototype speaker or environment, so that inter-speaker variances are minimized. For normalization both training and testing are performed with transformed speech. A widely used example for normalization is VTLN (Vocal Tract Length Normalization).

For performing the adaptation one starts from an initial trained system, and learns the transformation with an amount of new data, called **adaptation data**.

The adaptation process can be be

- **supervised**, which means that the word transcription (but usually not the alignment) of the adaptation data is known and used during the adaptation process, or
- **unsupervised**, where transcription is not known, but the recognition output of the initial system is used as the transcription. Certainly, unsupervised adaptation is the most desirable, but also most difficult. A poor initial recognition leads to a false transformation, so that there is a need of measures, how confident one can be in the recognition result of the initial system.

The adaptation mode can also be categorized as

- **off-line**, where the adaptation and test data are distinct and adaptation is performed completely before testing, and
- **on-line**, where adaptation is performed on the test data during testing,

or, also depending on the usage of the adaptation data, in

- **static** or **batch mode**, where all adaptation data is presented at once to the initial system for adaptation, or in
- **dynamic** or **incremental mode**, where the adaptation data is presented successively in small blocks and the system is refined after each block.

1.3 Existing Work

(1.3.1) Overview of Different Approaches

During the last years a lot of experiments have been done on adaptation, most of them for HMM-based speech recognition systems. Roughly speaking, one can categorize the following different approaches:

- Speaker categorization (clustering)

  Speakers or data are clustered and for every cluster a model is trained. The adaptation process consists of selecting the cluster that is most representative and use the corresponding model ([11, 13]). The method is simple but the drawbacks are that variations within clusters may be large and new data may not be well represented by any of the clusters. Gender dependent modelling is a special case of speaker clustering.

- Mapping of the feature space

  Input feature or model parameters are transformed to minimize differences between adaptation data and training data (for example minimize differences between speakers). This can be accomplished globally or phone specific. Experiments in [17, 20] show that linear or piecewise linear transformations perform good for HMM systems, which indicates that they capture well enough the differences being at the same time simple enough to generalize well.

- Reestimation of the parameter models

  Here the model parameters are transformed to improve the model accuracy for the adaptation data.

  It can be performed using Bayesian MAP (*Maximum a Posteriori*) iterative estimation, which uses a priori information about the distribution of the parameters ([10]). It has the advantage of converging asymptotically to a speaker dependent model as more adaptation data is available.

  A major problem for this approach is of course the estimation of the prior densities for the model, usually taken from a speaker-independent model. Just taking the speaker-independent models as initial parameters and not assuming any prior knowledge about the distribution of the parameters leads to the simpler (but poorer) ML (*Maximum Likelihood*) -estimation.

  Another problem with this technique is that it updates only parameters of models for which adaptation data exists. For limited adaptation data some models might not be observed in the example data and thus not been adapted. It has been shown that there is a relationship between different parameters of a speaker-independent system, so to overcome the problem of limited adaptation data it is possible to model this relationship and predict models for unobserved data using the updated models of the observed data. Linear regression relationships between parameters (learned from an SD-trained system) ([4]) and smoothing techniques ([14]) proved to be successful in a MAP-estimation framework, especially for very small amounts of adaptation data.

  A successful transformation-based approach to reestimate the parameter models is MLLR (*Maximum Likelihood Linear Regression*)([18]). The parameter models (usually only means of Gaussian distributions, sometimes variances too) are reestimated using regression-based transforms. The transformation parameters are determined such that the new model maximizes the adaptation data (EM-algorithm). To overcome the problem of limited data, transformations for different classes can share the same parameters.

(1.3.2) Reported Results for Adaptation of HMM Speech Recognition Systems

Some reported experimental results for adaptation of HMM-systems are shown in table 1 (supervised adaptation) and 2 (unsupervised adaptation).

The following conclusions can be drawn:

- The adaptation result is strongly dependent on the amount of training data. For supervised adaptation a boost is usually achieved with more than 4 minutes of speech. Especially ML-techniques require a

lot of data, for few data the performance can even degrade. For small amounts of adaptation data the effect of sparse data can be overcome by sharing of parameters, prediction or smoothing techniques.

- The adaptation result is highly dependent on the task and on the initial system.
- For unsupervised adaptation a high performance of the initial system is even more important. It has been reported that for an improvement a system better than 20 % WER is necessary ([23, 28]).

(1.3.3) Reported Results for Adaptation of Hybrid NN-HMM Speech Recognition Systems

For hybrid systems some results are presented in table 3 (supervised adaptation) and 4 (unsupervised adaptation).

We can observe that

- The adaptation result is highly dependent on the task, on the initial system, and on the amount of training data.
- LIN (Linear Input Network) gives good results for hybrid systems. It is similar to ML-techniques. Mixtures of LINs are slightly better and can be compared to MLLR - techniques.
- Retraining of the nets gave also promising results.
- Unsupervised training is also not consistent and depends on the performance of the initial system.

2 Confidence Measures

A **confidence measure** is a statistic which quantifies the degree of belief in the correctness of the result. A confidence measure of 1 indicates that we are sure the result is correct, 0 that we are sure it is incorrect.

There have been made a lot of investigations on confidence measures during the last two years. At the word level it has been reported that measures based on word lattices containing the n best hypotheses of the system proved to be most successful ([16, 21]), combined with other measures based on posterior probabilities, language model scores, number of occurences of words, speaking rate (durations), SNR, etc. ([15, 25, 8]).

An **acoustic confidence measure** is here defined as being based exclusively on the acoustical model, without taking into account higher-level sources of information like language model, semantic, etc. Acoustical confidence measures based just on local posterior probability estimations of the system were proposed and successfully tested for utterance rejection both for HMM-systems([22]) and for hybrid HMM-NN systems ([21]).

3 Implementation and Testing

3.1 System and Dataset

For all experiments the BRNN (Bidirectional Recurrent Neural Network) phoneme recognition system of Mike Schuster was used ([24]).

The BRNN estimates the posterior probabilities of the frame-class given the preprocessed acoustical input feature P(frame-class|frame). From these the scaled likelihoods of the frames are computed through the Bayes rule, and decoding is performed with the Viterbi algorithm.

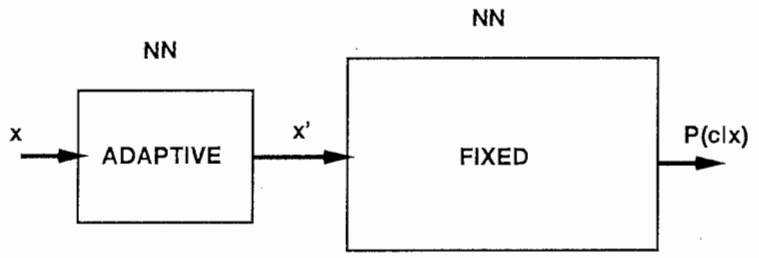
The BRNN used for experiments had 30 inputs (14 MFCC-coefficients, power and delta), 61 outputs (phonemes), and a total of 58211 weights.

The system was trained 64 iterations on the TIMIT (English continuous read speech) training set, using the cross-entropy objective function and the tanh activation function. For adapting and testing the TIMIT test set was used, containing 1344 utterances from 168 speakers (112 male, 56 female; 8 sentences per speaker) speaking 8 regional American dialects. The average length of the test set utterances is 2.8s.

The baseline performance of the speaker-independent system is 63% phoneme error rate.

3.2 Adaptation with a Linear Input Network

A linear input network was implemented for adaptation. The linear network has 930 weights (30 inputs make 30 * 30 + 30 weights) and is initialized as the identity transformation. Training is performed through backpropagation (RPROP) for the whole network (LIN + speaker-independent BRNN), but the weights of the BRNN are kept fixed, while just the weights of the LIN are updated (ML-training).



(a)

図 1: Architecture of LIN-BRNN system (LIN=ADAPTIVE, BRNN=FIXED)

The adaptation algorithm works as follows:

1. initialize 'LIN' as identity transformation
2. forward through 'LIN' and 'BRNN', calculate scaled likelihoods
3. Viterbi search to assign a target class for every frame
4. backpropagate error through 'LIN' and 'BRNN' and adjust weights in 'LIN'

For supervised adaptation step 3 is changed to a forced alignment of the already given phoneme sequence.

Experiments were performed both for supervised and unsupervised sentence, speaker and region adaptation, for all 168 speakers of the database.

3.3 Adaptation through Weight-Sharing

For adaptation with weight-sharing the weights of the speaker-independent system are clustered, for example with the kmeans algorithm using the Euclidean distance measure (other variants are possible as

options, but no experiments were performed). Than the BRNN is retrained with the adaptation data, so that all the weights in a cluster are updated equally, with averaging of the derivatives of the error for all the weights in the cluster.

Experiments were also performed both for supervised and unsupervised speaker as well as region adaptation, for all 168 speakers of the database.

3.4 Additional Confidence Measure for Unsupervised Adaptation

Unsupervised adaptation relies on a correct output of the system. For a wrong recognition output the adaptation can lead to a degradation of the performance instead to an improvement. Therefore it would be useful to adapt just to correctly aligned frames, ignoring the false aligned frames.

For that purpose we introduced an additional confidence measure at frame-level based on the local posterior probability estimation of the system and tested the results for unsupervised adaptation. The confidence measure used for every frame is the posterior probability of the recognized class for this frame as it is approximated by the BRNN.

$$confidence = p(\text{recognized class}|\text{input frame})$$

For every frame the backpropagated error at the output units is multiplied by the confidence.

Thus, if the confidence is high (close to 1) it means that the system is quite sure that the result is correct, the whole error at the output unit is backpropagated and the net learns to recognize this frame even better. If the confidence is small (close to 0), the frame is not used for adaptation.

The linear correlation coefficient of *confidence* with known correct/false assignment for frames is 0.48 for a test set of 10 speakers (80 sentences) using the SI BRNN-system.

3.5 Experimental Results

Supervised adaptation has been performed in off-line batch mode (separate adaptation and test data presented at once to the system) and unsupervised adaptation in on-line batch mode (adaptation and test data the same, presented at once to the system) for sentences, speakers and regions. The best results for different experiments are listed in tables 5 and 6.

There have been done different experiments for determining the values for RPROP-parameters for both adaptation with LIN and through retraining with weightsharing (dependent on the number of shared weights). Only the best results are listed here. Also there have been done experiments for determining the number of adaptation iterations and dependency on amount of adaptation data.

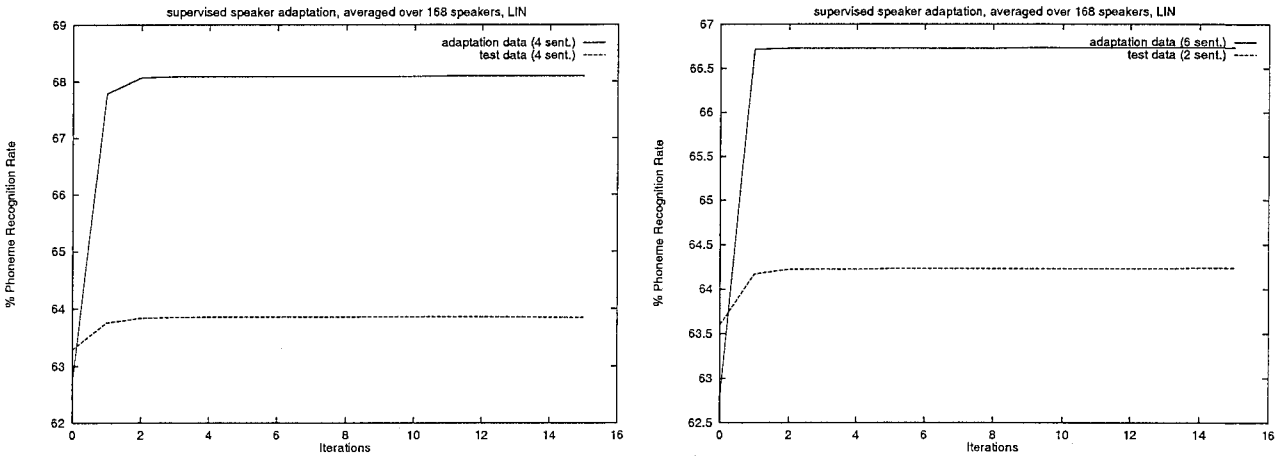
The learning curves for 15 iterations of supervised adaptation for all 168 speakers are listed in Fig. 2, 3, and 4.

Fig. 2 shows that the system doesn't improve on the adaptation data after 1-3 iterations for 4 sentences adaptation data and after 1-2 iterations for 6 sentences adaptation data, which implies that the transformation is too simple (or that the number of parameters must be increased - which is not possible for a LIN). The effect is even worse for region adaptation (between 40 and 120 sentences/region), where performance drops (Fig. 3).

With a BRNN (nonlinear transformation) the performance improves continually on the adaptation data, but due to the small number of sentences (4 adaptation data/speaker) the net doesn't generalize. A net with 930 weight-clusters (same number as the LIN-parameters) is compared to a net with 2500 weights in Fig. 4. The net with 2500 weights adapts more accurate to the adaptation data due to more parameters, but has worse results on the testing data, since it overlearns the adaptation data.

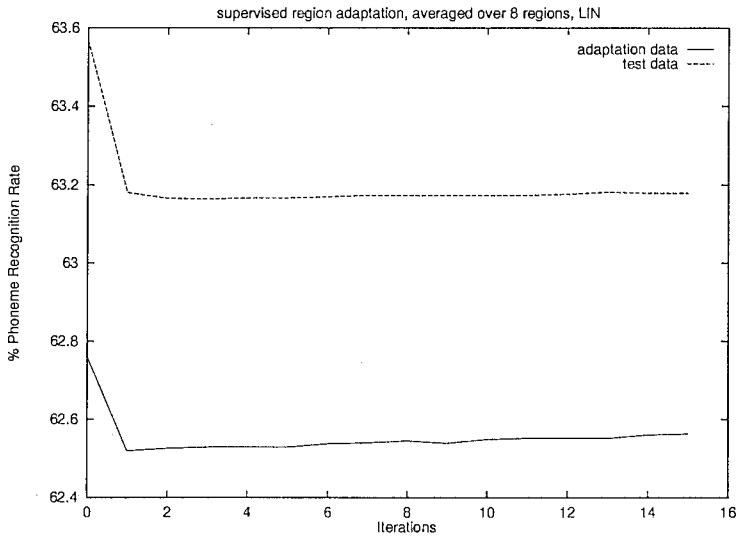
For unsupervised adaptation (Fig. 5, 6, 7, 8) in average the performance degrades, which confirms other reports on similar experiments. This is due to the fact that the speaker-independent system has only 63% phoneme recognition rate, with more than one third of the alignment being false, such that the system doesn't have any mechanism to avoid adaptation to the wrongly aligned frames.

Introducing a simple acoustical confidence measure based on posterior probabilities the results for unsupervised adaptation after one iteration improved with 0.3 % relative for sentence and speaker adaptation



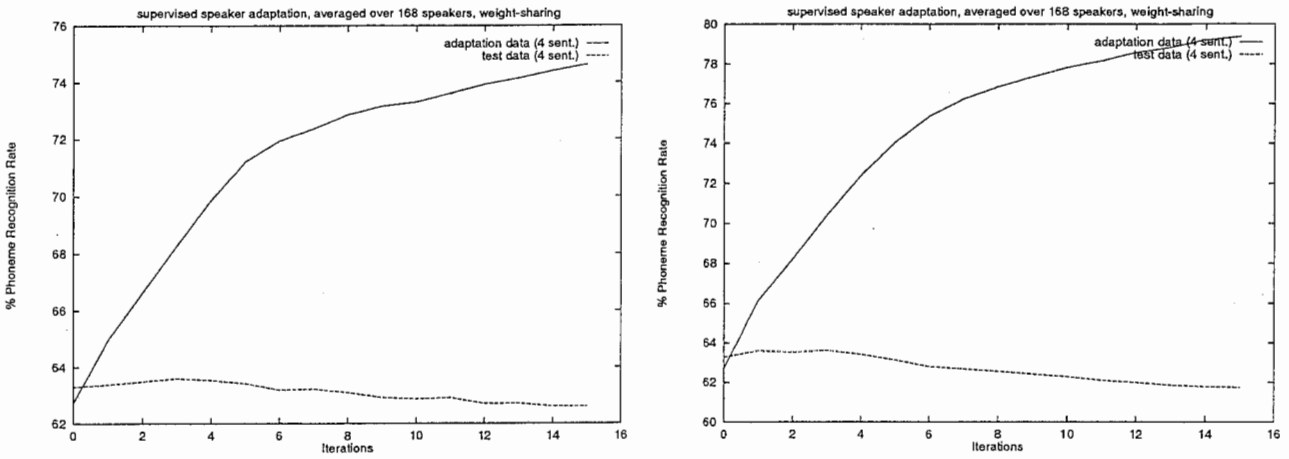
(a) (b)

図 2: Supervised speaker adaptation with LIN with (a) 4 adaptation sentences, 4 test sentences (b) 6 adaptation sentences, 2 test sentences for every speaker



(a)

図 3: Supervised region adaptation with LIN. The sentences for every speaker of the same regions has been divided in half adaptation half test data



(a) (b)

図 4: Supervised speaker adaptation with BRNN through retraining with weight-sharing for (a) 930 weights (b) 2500 weights

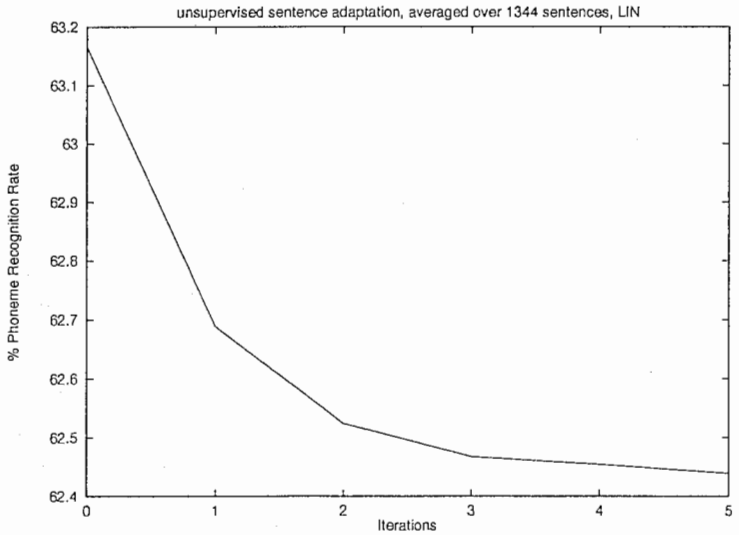
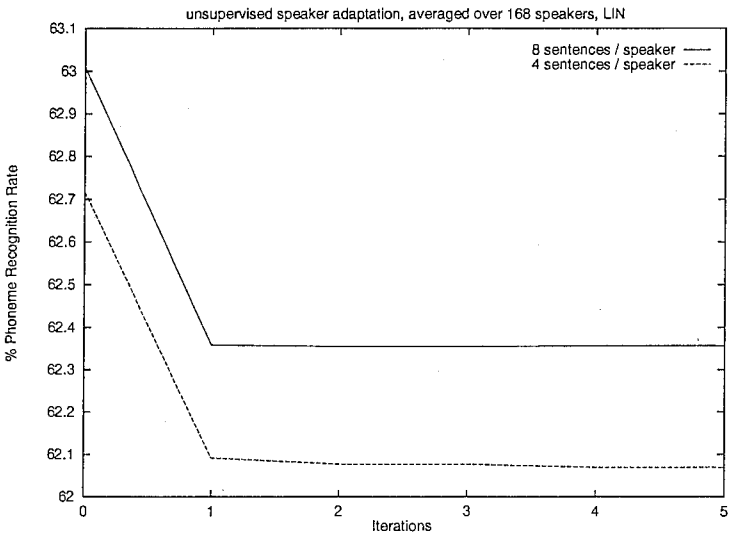
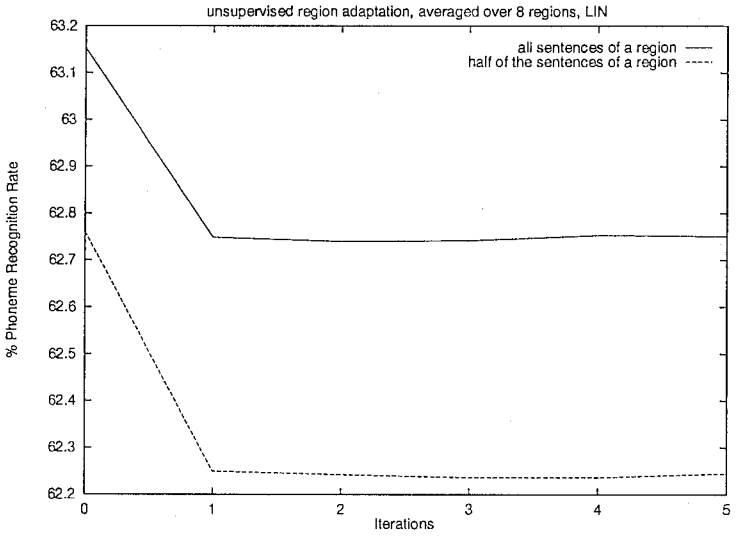


図 5: Unsupervised adaptation with LIN for sentences



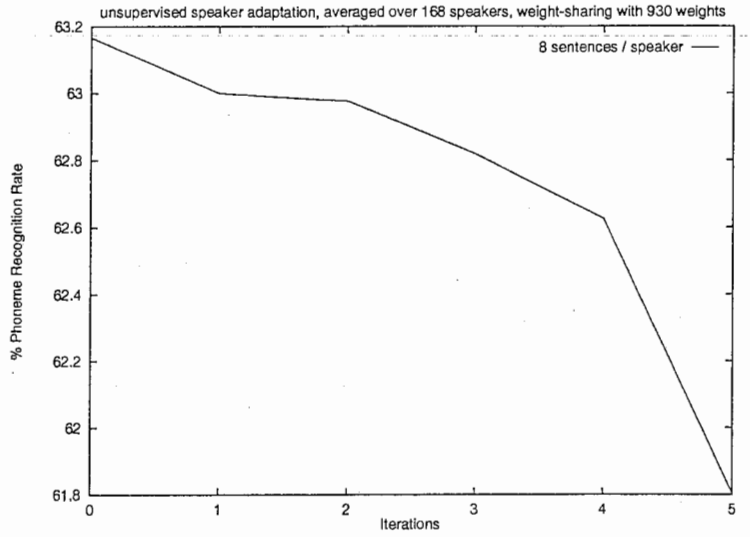
(a)

図 6: Unsupervised adaptation with LIN for speakers



(a)

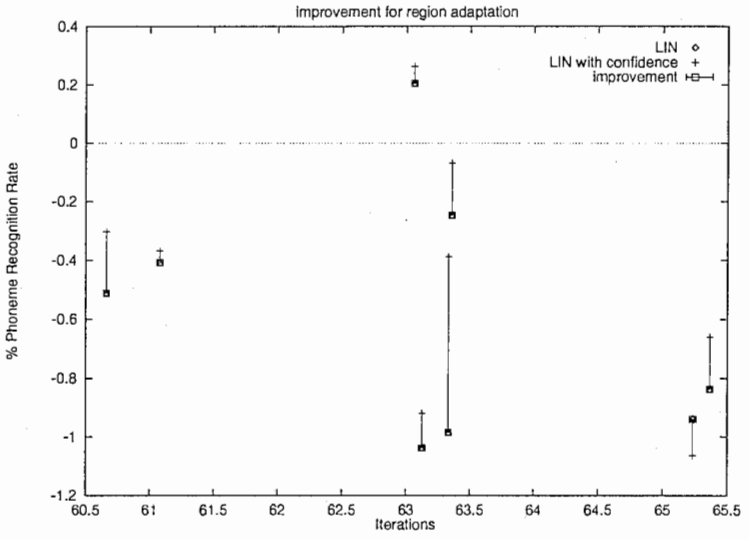
図 7: Unsupervised adaptation with LIN for regions



(a)

図 8: Unsupervised speaker adaptation with retraining through weight-sharing

and with 1.8 % relative for region adaptation in comparison with the case without confidence measure (see table 6). The relative improvements as a function of recognition rate are plotted for the LIN case without versus with confidence for region adaptation in Fig. 9, we can see that 7 out of 8 regions get better results, just one region decreased.



(a)

図 9: Unsupervised region adaptation improvements dependent on initial phoneme recognition rate, with vs. without confidence measure

4 Conclusion

In this work we have implemented and tested three methods of adaptation for a hybrid BRNN-HMM phoneme recognition system on the TIMIT task.

The linear transformation of the input feature with a LIN (Linear Input Network) is a simple adaptation method which gave promising results in previous reported work. It seems that for our task and system it is too simple and has too few parameters (the number of parameters is fixed), so that a nonlinear transformation or a mixture of LINs might give better results.

Retraining of the BRNN with weight-sharing is a method with which one can control the number of parameters to be adapted. The learned transformation is more complex, but supervised adaptation gave only small improvements on testing data due to the limited amount of adaptation data.

For unsupervised adaptation the initial system is too weak (63% phoneme accuracy) for giving consistent improvements. But introducing only a simple acoustical confidence measure for every frame based on posterior probabilities of the recognized class we could consistently improve the result for unsupervised adaptation. This makes us think that a more sophisticated confidence measure could give an even better improvement.

参考文献

[1] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist Speaker Normalization and Adaptation", in *Proc. European Conference on Speech Communication and Technology*, pp. 2183-2186, 1997.

[2] V. Abrash, "Mixture Input Transformations for Adaptation of Hybrid Connectionist Speech Recognizers", in *Proc. European Conference on Speech Communication and Technology*, pp. 299-302, 1997.

[3] M. Afify, Y. Gong, and J.-Paul Haton, "Correlation Based Predictive Adaptation of Hidden Markov Models", in *Proc. European Conference on Speech Communication and Technology*, pp. 2059-2062, 1997.

[4] S. M. Ahadi-Sarkani and P. Woodland, "Rapid Speaker Adaptation Using Model Prediction", in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 684-687, 1995.

[5] S. M. Ahadi-Sarkani, *Bayesian and Predictive Techniques for Speaker Adaptation*, PhD Thesis, University of Cambridge, January 1996.

[6] X. Aubert and E. Thelen, "Speaker Adaptive Training Applied to Continuous Mixture Density Modeling", in *Proc. European Conference on Speech Communication and Technology*, pp. 1851-1854, 1997.

[7] C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, 1995.

[8] L. Chase, "Word and Acoustic Confidence Annotation for Large Vocabulary Speech Recognition", in *Proc. European Conference on Speech Communication and Technology*, pp. 815-818, 1997.

[9] V. Digalakis, "On-line Adaptation of Hidden Markov Models Using Incremental Estimation Algorithms", in *Proc. European Conference on Speech Communication and Technology*, pp. 1859-1862, 1997.

[10] R. Duda and P. Hart, *Pattern Recognition and Scene Analysis*, John Wiley and Sons, 1973.

[11] G. Yuqing and P. Mukund and M. Picheny, "Speaker Adaptation Based on Pre-Clustering Training Speakers", in *Proc. European Conference on Speech Communication and Technology*, pp. 2091-2094, 1997.

[12] G. Williams and S. Renals, "Confidence Measures for Hybrid HMM/ANN Speech Recognition", in *Proc. European Conference on Speech Communication and Technology*, pp. 1955-1958, 1997.

[13] T. J. Hazen and J. R. Glass, "A Comparison of Novel Techniques for Instantaneous Speaker Adaptation", in *Proc. European Conference on Speech Communication and Technology*, pp. 2047-2050, 1997.

[14] Q. Huo and C.-H. Lee, "Combined On-line Model Adaptation and Bayesian Predictive Classification for Robust Speech Recognition", in *Proc. European Conference on Speech Communication and Technology*, pp. 1847-1850, 1997.

[15] T. Kemp and T. Schaaf, "Confidence measures for spontaneous speech", in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 875-878, 1997.

[16] T. Kemp and T. Schaaf, "Estimating Confidence Using Word Lattices", in *Proc. European Conference on Speech Communication and Technology*, pp. 827-830, 1997.

[17] C. J. Leggetter, *Improved Acoustic Modeling for HMMs Using Linear Transformations*, PhD Thesis, Cambridge University Engineering Departement, February 1995.

[18] C. J.Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", in *Computer Speech and Language*, 9(2), pp. 171-185, April 1995.

[19] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker Adaptation for Hybrid HMM-ANN Continuous Speech Recognition Systems", in *Proc. European Conference on Speech Communication and Technology*, pp. 2171-2174, 1995.

[20] L. Neumeyer, A. Sankar, and V. Digalakis, "A Comparative Study of Speaker Adaptation", in *Proc. European Conference on Speech Communication and Technology*, pp. 1127-1130, 1995.

[21] G. Williams and S. Renals, "Confidence Measures for Hybrid HMM/ANN Speech Recognition", in *Proc. European Conference on Speech Communication and Technology*, pp. 1955-1958, 1997.

[22] Z. Rivlin, M. Cohen, V. Abrash, and T. Chung, "A Phone-Dependent Confidence Measure for Utterance Rejection", in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 515-518, 1996.

[23] D. Rtischev, Q. Huo, and H. Singer, *Implementation and Testing of Quasi-Bayes Speaker Adaptation Algorithm*, ATR Technical Report 1996.

[24] M. Schuster, "Bidirectional Recurrent Neural Networks in Speech Recognition", in *Proc. of the Acoustical Society of Japan Meeting*, Okayama, 1996.

[25] M.-H. Siu, H. Gish, and F. Richardson, "Improved Estimation,Evaluation and Applications of Confidence Measures for Speech Recognition", in *Proc. European Conference on Speech Communication and Technology*, pp. 831-834, 1997.

[26] Steve Waterhouse, Dan Kershaw, and Tony Robinson, "Smoothed Local Adaptation of Connectionist Systems", in *Proc. Intern. Conf. on Spoken Language Processing*, pp. 1313-1316, 1996.

[27] R. L. Watrous, "Speaker Normalization and Adaptation Using Second-Order Connectionist Networks", in *IEEE Transactions on Neural Networks*, Vol.4, No. 1, January 1993.

[28] Puming Zhan, Martin Westphal, Michael Finke and Alex Waibel, "Speaker Normalization and Speaker Adaptation - a Combination for Conversational Speech Recognition", in *Proc. European Conference on Speech Communication and Technology*, pp. 2087-2090, 1997.

System	Database	Method	Adaptation data	Test data	SI	Adapted	Rel. impr.
HTK ([5])	ARPA RM1 (cont.)	MAP	1 sent.(3sec)/sp. (12 sp.)	100 sent./sp.	5.92	5.86	1.0
			5 sent.(15sec)/sp. (12 sp.)	"	5.92	5.62	5.0
			600 sent.(1800sec)/sp. (12 sp.)	"	5.92	2.38	59.7
		ML	1 sent.(3sec)/sp. (12 sp.)	"	5.92	39.76	-571.6
			5 sent.(15sec)/sp. (12 sp.)	"	5.92	63.05	-965.0
			600 sent.(1800sec)/sp. (12 sp.)	"	5.92	2.36	60.1
		RMP	1 sent.(3sec)/sp. (12 sp.)	"	5.92	5.44	8.1
			10sent. (30sec)/sp. (12 sp.)	"	5.92	5.19	12.3
			600 sent.(1800sec)/sp. (12 sp.)	"	5.92	2.34	60.2
HTK ([5])	native 1994 WSJ Spoke S3	MAP	40 sent./sp. (284sp.)	20 sent./sp.	5.58	5.20	6.8
		RMP	40 sent./ sp.	20 sent./sp.	5.58	5.03	9.9
	nonnative 1994 WSJ Spoke S3	MAP	40 sent./sp.	20 sent./sp.	19.59	16.56	15.5
		RMP	40 sent./ sp.	20 sent./sp.	19.59	16.37	16.4
SRI's DECIPHER ([20])	native 1994 WSJ S3	ML + Bayes	40 sent./sp. (10 sp.)	20 sent./sp.	20.9	17.5	16.2
	nonnative 1994 WSJ S3	ML + Bayes	20 sent./sp. (5sp.)	20 sent./sp.	24.9	14.4	42.1
			40 sent./sp. (11 sp.)	20 sent./sp.	23.1	10.5	54.7
ATR ([23])	ATR SSD (spont.)	Quasi-Bayes (MAP)	15s/sp. (3sp.)		19.7	18.8	4.56
			260s/sp. (3sp.)		19.7	13,4	31.9
		MAP + VFS	15s/sp. (3sp.)		19.7	17.6	10.65
			260s/sp. (3sp.)		19.7	13,9	29.44
JANUS ([28])	SSST	MLLR	(9sp.)		21.8	15.3	12.5
HTK ([3])	TIMIT	MAP	2 sent./ sp. (24 sp.)	?	52.27	49.60	5.1

Table 1: Overview of results (% WER) for supervised adaptation of HMM-systems (sent.= sentence, sp.= speaker)

System	Database	Method	Adaptation data	SI	Adapted	Rel. impr.
HTK ([5])	ARPA RM1 (cont.)	MAP	10 sent.(30s)/sp. (12 sp.)	5.92	5.9	0.5
			600 sent.(1800s)/sp.	5.92	4.69	20.8
JANUS ([28])	SSST (spont.)	MLLR	(9sp.)	21.8	21.3	2.2
([6])	WSJ S0	SAT	1sent./sp. (20 speakers)	7.93	6.28	20.7
SRI's DECIPHER ([9])	WSJ S3	incremental ML	20 sent./sp.	27.4	19.6	28.4

Table 2: Overview of results (% WER) for unsupervised adaptation of HMM-systems

System	Database	Method	Adaptation data	Test data	SI	Adapted	Rel. impr.
SRI's DECIPHER ([1])	male native WSJ S3 (read)	LIN	30+10 sent./sp.(4 sp.)	40 sent./sp.	21.77	17.26	15.3
		retraining of MLP	"	"	21.77	18.44	15.2
		LIN + retrain MLP	"	"	21.77	16.91	18.7
	male nonnative WSJ S3 (read)	LIN	30+10 sent./sp.(5 sp.)	40 sent./sp.	24.5	19.2	21.63
		retraining of MLP	"	"	24.5	15.7	35.91
		LIN + retrain MLP	"	"	24.5	15.6	36.32
SRI's DECIPHER ([2])	male nonnative WSJ S3 (read)	mixtures of LIN's	30+10 sent./sp.(5 sp.)	40 sent./sp.	33.6	22.5	33.2
MLP-HMM ([19])	DARPA RM1 (read)	LIN	80 sent./sp.(12 sp.)	20 sent./sp.	8.5	5.5	35.29
		retrain MLP	"	"	8.5	6.2	27.05
RNN-HMM ([19])	DARPA RM1 (read)	LIN	80 sent./sp.(12 sp.)	20 sent./sp.	8.4	6.4	23.8
		"	600 sent./sp.(12 sp.)	100 sent./sp.	8.4	3.9	53.5
		retrain RNN	"	"	8.4	6.9	17.85

Table 3: Overview of results (% WER) for supervised adaptation of NN-HMM-systems

System	Database	Method	Adaptation data	SI	Adapted	Rel. impr.
ABBOT (RNN-HMM) ([26])	1995 ARPA Hub 3 MUM (clean speech, multiple microphones)	LIN	15 sent.sp. (20sp.)	18.5	15.9	12.1
		mixtures of LINs(MLLR?)	15 sent.sp. (20sp.)	18.5	15.7	13.2

Table 4: Overview of results (% WER) for unsupervised adaptation of NN-HMM-systems

Method	Task	Adaptation data	Test data	SI	Adapted	Rel. impr.
LIN	speaker adaptation	4 sent.(12s)/speaker (168 sp.)	4 sent.(12s)/speaker	63.28	63.86	0.9
		6 sent.(18s)/speaker (168 sp.)	2 sent.(6s)/speaker (168 sp.)	63.60	64.23	1.1
	region adaptation	44-104sent.(132-312s)/region (8 regions)	44-104sent.(132-312s)/region (8 regions)	63.56	63.18	-0.5
WS-500 weights	speaker adaptation	4 sent.(12s)/speaker (168 sp.)	4 sent.(12s)/speaker	63.28	63.79	
WS-930 weights	"	"	"	63.28	63.66	
WS-2500 weights	"	"	"	63.28	63.62	

Table 5: Overview of results (% WER) for supervised adaptation of BRNN-HMM-systems

Method	Task	Adaptation data	SI	Adapted	Rel. impr.
LIN	sentence adaptation	3s/sentence	63.16	62.68	-0.6
	speaker adaptation	8 sent.(24s)/speaker (168 sp.)	63.00	62.35	-1.0
	region adaptation	88-208sent.(264-624s)/region (8 regions)	62.15	62.75	0.9
LIN + confidence	sentence adaptation	3s/sentence	63.16	62.95	-0.3
	speaker adaptation	8 sent.(24s)/speaker (168 sp.)	63.00	62.55	-0.7
	region adaptation	88-208sent.(264-624s)/region (8 regions)	62.15	63.71	2.4
WS-500 weights	speaker adaptation	8 sent.(24s)/speaker (12 sp.)	64.23	65.23	1.5
WS-930 weights	"	"	64.23	64.78	0.7
WS-2500 weights	"	"	64.23	64.60	0.6

Table 6: Overview of results (% WER) for unsupervised adaptation of BRNN-HMM-systems