Knowledge-Based and Automated Clustering in MLLR Adaptation of Acoustic Models for LVCSR

Michal Borský, Petr Pollák
Czech Technical University in Prague
Faculty of Electrical Engineering
K13131 CTU FEE, Technická 2, 166 27 Prague 6, Czech Republic
Email: {borskmic,pollak}@fel.cvut.cz

***Abstract*—This paper describes the analysis of the performance of MLLR-based speaker adaptation in a large vocabulary continuous speech recognition system. Two different approaches of clustering in MLLR-adaptation with more regression classes, knowledge-based clustering and automatic clustering were analysed. The contribution of mentioned acoustic model adaptation using these two clustering approaches were compared based on the word error rate ratio (WERR) of target LVCSR. Realized study proved that the knowledge-based clustering may bring improvement comparable to the tree-based clustering, when only a few transformation classes are manually defined.**

***Keywords*—MLLR, LVCSR, adaptation, regression classes, acoustic modelling**

I. Introduction

The development of automatic speech recognition (ASR) in the last few decades has resulted in the emergence of many ASR systems in our everyday life. The results achieved in the field have started a transition from isolated word recognition suitable for controlling simple devices to a continuous speech recognition, where the transcription of a much larger and complex text is required. Possible applications are automatic transcriptions of spontaneous speech in real-time [1], [2], converting the audio recordings into text, speech controlled devices used in cars or households and many others.

The two core parts of any large vocabulary continuous speech recognition (LVCSR) system with crucial influence on the accuracy are acoustic models (AM), for modelling the acoustic part of the speech, and language models (LM). The data required for proper AM training is enormous. Possible solution to this problem is the use of audio data from many different speakers, which results in creating speaker-independent (SI) AM. To circumvent this aspect and improve the performance of the whole system practically all of the state-of-the-art LVCSR systems employ subsequent AM adaptation [3], [4] to fit the AM on the unique characteristics of any speaker or environment conditions.

Nowadays, the two most frequently used adaptation techniques are Maximum-Likelihood Linear Regression adaptation (MLLR) and Maximum A Posteriori (MAP) [5] adaptation, when the former yield better results when little adaptation data is present. This particular trait is achieved due to applying the same transformation for components which are perceived as similar and are therefore tied to the same transformation class. For this reason the performance of the MLLR adaptation is dependent not only the quality and quantity of adaptation data but on the initial clustering of components into the classes as well.

In this paper we investigate the effect of automatic clustering on the performance of the LVCSR system and compare it to the clustering based on phone creation and articulation, in another words phonetics. A similar research [6] showed that knowledge-based clustering with only two defined classes, one for speech models and for non-speech models, may bring considerable improvement to accuracy in real conditions, where a high level of background noise was present and thus the adaptation was focused more on the environment than on the speaker. In our research the attraction is put on speaker adaptation.

The paper is organized as follows: the MLLR speaker adaptation with a special focus on two discussed clustering methods are reviewed in the Section II. The Section III describes the experimental setup. In the Section IV the achieved results are presented and the Section V concludes with a final discussion.

II. MLLR adaptation

The initial solution to speaker adaptation when little adaptation data is available by clustering similar components into classes was described in [8]. The principle is to maximize the likelihood of the adaptation data, when an affine transform for the Gaussian parameters of the SI models is computed. In most cases the MLLR adaptation is applied only to the means of distributions, meaning that the regression matrices are tied only across the Gaussian means and the variances are left unchanged. For the adaptation to be effective the regression matrices are to be tied across components which will share the same transform. Since this information is not available prior to computing the transformations, a beforehand estimation of classes compositions has to be made. Two different approaches

to this problem are available: automatically tie the components with similar parameters or use information about phonetic classes.

A. Automatic clustering

The automatic clustering of Gaussian components into classes is based on the assumption that similar components should share the same transformation. Therefore the automatic clustering is based on the proximity in acoustic space and no distinction as to which HMM the particular component belongs is taken into account. The resulting composition of classes is done according to information (stastics) obtained at training phase of AM creation. This means that the error from this phase quantifiable as the WER transforms into a possible error for clustering as well. The final number of classes is determined by the amount of adaptation data, when in the most simple case only a single class is determined and all components share the same transform. The number of classes increases as more adaptation data becomes available which allows for more precise adaptation.

B. Knowledge-based clustering

The knowledge based clustering is based on the information about phone creation. In this case the phones with similar articulation and vocal tract characterization during their utterance are assumed to be acoustically similar and should therefore share the same regression class and transform. For that reason the distribution of phones across the regression classes corresponds to their separation into phonetic classes defined by phonetics.

In our study the phones division was done according to basic phonetic categorization, when for the Czech language there are 10 vowels, 28 consonants, 3 diphthongs and allophones, altogether creating a set of generally recognized 44-46 phones [11]. This division was too coarse and further improvement was desired, so and extra division of consonants into 6 standard classes was applied: vowels, nasals, liquids, fricatives, plosives and affricates. A special class was added for the diphthongs (”au, ”ou”, ”eu”) and ”silence”.

The acoustic modelling was done at the level of context-dependent phones. As the number of triphones was too large to manually determine the eventual classes compositions a further simplification at this stage of the work was used. It was assumed that the phonetic representation of a triphone is mostly determined by its middle monophone [9] and all the triphones with the same middle monophone were clustered into a same class. Finally the same clustering policy was adopted to the mixtures and states as well, when for all mixtures and emitting states for a particular triphone the same linear transformation was computed. The described strategy of distributing triphones across classes resulted in a fairly straightforward composition presented in Tab. I, with the total number of 8 expertly determined classes.

This basic setup was expanded later on when in the next step $class2$, which contained both the long and short versions of the vowels, was split into two classes following the lip position, in Tab. II, or tongue movement in Tab. III. These tables show only the divided vowels class as the rest of the classes remained the same.

TABLE I
KNOWLEDGE-BASED TRIPHONE CLASSES

Class	*Phone*
Class 1	silence
Class 2	(*-a+*),(*-á+*),(*-e+*),(*-é+*),(*-i+*),(*-í+*), (*-o+*),(*-ó+*),(*-u+*),(*-ú+*)
Class 3	(*-au+*),(*-eu+*),(*-ou+*)
Class 4	(*-l+*),(*-r+*),(*-j+*)
Class 5	(*-m+*),(*-n+*),(*-nn+*), (*-ng+*),(*-mv+*)
Class 6	(*-f+*),(*-v+*),(*-s+*),(*-z+*),(*-ss+*),(*-zz+*), (*-ch+*),(*-h+*),(*-rr+*),(*-rrr+*)
Class 7	(*-p+*),(*-b+*),(*-t+*),(*-d+*),(*-tt+*),(*-dd+*), (*-k+*),(*-g+*)
Class 8	(*-c+*),(*-dz+*),(*-cc+*),(*-dzz+*)

TABLE II
VOWELS CLASS DIVISION - A) LIP POSITION

Class	*Phone*
Class 2.1a	(*-a+*),(*-á+*),(*-e+*),(*-é+*),(*-i+*),(*-í+*),
Class 2.2a	(*-o+*),(*-ó+*),(*-u+*),(*-ú+*)

TABLE III
VOWELS CLASS DIVISION B) TONGUE MOVEMENT

Class	*Phone*
Class 2.1b	(*-e+*),(*-é+*),(*-i+*),(*-í+*),
Class 2.2b	(*-a+*),(*-á+*), (*-o+*),(*-ó+*),(*-u+*),(*-ú+*)

III. EXPERIMENTAL SETUP

In this section the setup for evaluating proposed clustering methods is described. This includes the AM creation, datasets description and performance evaluation.

A. LVCSR setup

The MLLR clustering methods were evaluated in a LVCSR system created for our experiments using the HTK Toolkit [10], when for feature extraction the following configuration was used : 13 MFCCs complemented by their 1st and 2nd derivatives, the frame length of 25ms and 10ms frame shift, Hamming window and liftering of order 22.

The acoustic modelling was done at first on the set of 43 Czech monophones, complemented by a model for silence and short-pause (sp) with a standard left-to-right 3 state structure, no state skips allowed except for the ”sp” model. This monophone set was expanded into a set of cross-word triphones and then statistically state-tied and compacted from the initial number of 83k models into a final group containing 15k context dependent HMMs, adding 6 mixtures per state and no stream-splitting. The speaker independent (SI) models were trained using the flat-start procedure.

For a language model an internal trigram LM developed for the LVCSR testing in Czech language [7] was used. This model was created from the SYN2006PUB 5-gram Corpus, using the Kneser-Nay discounting with a vocabulary size of 340k. The n-gram statistics for this

corpus are available under license agreement and were acquired entirely from newspaper text [13].

For decoding a publicly available decoder HDecode was used with following parameters: pruning beam factor was set to 200, word insertion penalty to -10 and LM weight factor to 10.

B. Data for Experiments

The speech recognition and adaptation experiments were carried out with data from Czech SPEECON database [12]. This database contained audio data recorded under different kinds of environment conditions from 580 speakers with 300 utterances per speaker, varying in the level of the background noise and using multiple microphones. For this experiment the whole SPEECON was carefully revised and appropriate disjunct training, adaptation and testing sets were extracted.

The training set contained 60k utterances from 190 different speakers of an overall length of 51 hours, using only the audio data of a high-quality with rich phonetic content, i.e. data recorded with a headset microphone in a clean OFFICE environment with low level of background noise.

The performance of the baseline system and clustering approaches were tested out on the set of 275 utterances containing only the whole sentences with 27.5 minutes of overall length. This set contained 11 different adult speakers of both genders with various speech accents corresponding to their place of origin (Moravia, Silesia, Bohemia). The utterances in this set were selected due to their sentence structure, which made them suitable for a LVCSR testing. For each speaker approximately 23 utterances were available, recorded in the same conditions as was used for the training and adaptation task, thus minimizing the impact of a different environment or microphone.

The adaptation was estimated using the set of approx. 170 utterances for each speaker of all overall length of nearly 4 minutes. The set contained mostly utterances of a single word or a phrase, e.g. a name, address pronunciation or a simple command used in speech device control.

C. Performance evaluation

The results of speech recognition were evaluated using standard criteria of (*WER*) defined as :

$$WER = \frac{S + D + I}{N}.100\% \qquad (1)$$

where N was the total number of words and *S, D, I* represented the number of substituted, deleted and inserted words respectively. The contribution of adaptation was evaluated in terms of Word Error Rate Reduction (*WERR*) computed again the baseline system *WER* and defined as :

$$WERR = \frac{WER_{base} - WER_{adapt}}{WER_{base}}.100\% \qquad (2)$$

IV. Results

Experimental results in this section compare the impact of automatic and knowledge-based clustering on the performance of the LVCSR system under real conditions.

The adaptation process started with training the SI model, for which the regression classes were determined using the strategies described in Section 2. For both clustering methods the adaptation was carried out in two steps. At first all components were tied into a single class and a global transform was computed. For automatic clustering the decision tree was constructed using the centroid splitting algorithm when the effect of multiple number of classes was tested out, starting with 2 and increasing in each step until the WERR stopped increasing. Our goal was to map the effect of the number of classes on the final accuracy. The knowledge-based clustering was initiated with a global transform and continued with transforms for manually defined classes presented in Tab. I, Tab. II and Tab. III.

A. Results with knowledge-based clustering

The results for knowledge-based clustering is summarized in Tab. IV, when the values labeled as $Setup\_1$ corresponds to clustering presented in Tab. I, $Setup\_2$ to Tab. II and $Setup\_3$ to Tab. III. The best overall results were achieved using the basic $Setup\_1$, when the reduction of 5.69% in $mean(WER)$ was observed. Both $Setup\_2$ and $Setup\_3$ yielded very similar results to each other with 5.35% and 5.26% of reduction in $mean(WER)$. This shows that for the given amount of adaptation data there was no clear distinction as to which clustering was used.

TABLE IV
Results for Knowledge-based Clustering

	WER [%]			
$Speaker\_ID$	$Baseline$	$Setup\_1$	$Setup\_2$	$Setup\_3$
066	29.31	26.29	26.72	27.16
108	22.07	14.41	15.77	16.67
110	14.73	13.83	13.84	12.95
379	16.08	13.99	12.59	13.29
430	15.82	13.7	13.78	13.78
432	20.88	13.19	13.19	13.19
475	31.87	22.31	21.91	21.91
485	44.57	33.7	35.33	35.33
486	26.06	21.81	21.28	21.28
487	28.39	26.27	27.97	26.69
488	32.64	20.21	21.24	22.28
$mean(WER)$	25.67	19.98	20.32	20.41
$mean(WERR)$		20.91	20.14	19.91

B. Results with Automated Clustering

Tab. V summarizes the results for automatic clustering with multiple classes, when for 8 regression classes the reduction of 5.59% in $mean(WER)$ was observed, which corresponds to 20.25% in $WERR$. This result was also the best. After a slight decline for 12 classes, the values of WER began to rise again, but 32 classes proved to be a threshold, since no improvement past this number was measured. Highest

TABLE V
RESULTS FOR AUTOMATIC CLUSTERING

	WER [%]						
$Speaker\_ID$	$Baseline$	$Classes\_2$	$Classes\_4$	$Classes\_8$	$Classes\_12$	$Classes\_16$	$Classes\_32$
066	29.31	26.29	25.86	27.16	27.59	28.02	26.29
108	22.07	18.02	15.32	14.86	14.86	15.32	15.32
110	14.73	13.39	12.95	12.95	13.84	13.84	13.84
379	16.08	12.59	12.59	14.69	13.99	11.89	12.59
430	15.82	15.31	15.82	14.8	14.29	14.29	13.78
432	20.88	14.29	14.19	14.19	13.19	13.19	13.74
475	31.87	25.5	22.71	21.91	24.7	20.72	20.31
485	44.57	33.15	34.24	33.7	35.33	36.41	34.78
486	26.06	20.21	20.74	20.21	21.28	22.34	21.81
487	28.39	31.36	28.39	26.29	24.58	28.39	28.81
488	32.64	22.28	20.73	20.21	21.76	21.76	23.32
$mean(WER)$	25.67	22.12	20.32	20.08	20.49	20.56	20.41
$mean(WERR)$		16.68	19.72	20.25	19.3	19.49	19.78

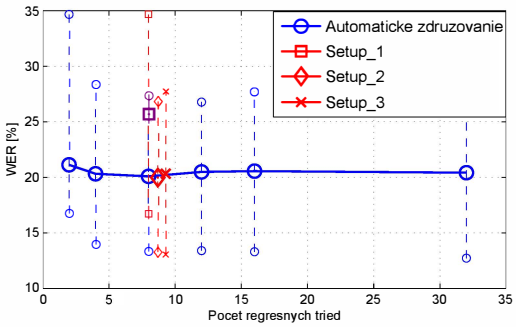

Fig. 1. Mean values of WERR for Automatic and Knowledge-based Clustering

reduction in WER was measured for speakers with relatively high WER before adaptation and likewise speakers with relatively low baseline WER showed only small values of $WERR$.

V. CONCLUSION

The aim of this paper was to analyse the effect of two different clustering approaches on the performance of the LVCSR system. The first clustering method was based on the grouping components with similar acoustic space position and the second method on commonly recognized phonetic classes.

Fig. 1 shows that the best results were achieved using 8 regression classes for both the automatic and knowledge-based clustering. Also that both methods yield very similar results of improvement, with 20.25% and 20.91% of $WERR$ respectively. One thing to note is the high value of variance of WER for the baseline system, when the difference between the best and the worst Speaker (Speaker_110 and Speaker_485) was 29.84%.

The amount of adaptation data is one of the key elements since the final number of classes for the automatic approach is determined appropriately to this amount. This is not the case for manual division, where the number was based solely on phonetics and would had to be revised if the amount proved to be insufficient. This is seen as major drawback of manual clustering.

ACKNOWLEDGMENT

Research described in the paper was supported by internal CTU grant SGS12/143/OHK3/2T/13 Algorithms and Hardware Realizations of Digital Signal Processing.

REFERENCES

[1] Nouza,J., Zdansky,J.,David,P. Fully automated approach to broadcast news transcription in Czech Language. *Text, Speech and Dialogue: Lecture Notes in Computer Science*, 2004, vol. 3206/2004, p.401-408.5

[2] Psutka,J.,Psutka,J.,Ircing,P.,Hoidekr,J. Recognition of spontaneously pronounced TV ice-hokey commentary. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokyo (Japan), 2003 p. 83-86.

[3] Cerva,P., Nouza,J., Kolerenc, J., David, P.Improved Transcription of Czech Parliamentary Speeches by Acoustic and Language Model Adaptation. In *SPEECOM'2006* St. Petersburg, 2006 p.25-29.

[4] Cerva, P., Zdansky, J., Silovsky, J., Nouza, J. Study on Speaker Adaptation Methods in the Broadcast News Transcription Task. In *Lecture Notes in Artificial Intelligence, Text, Speech and Dialogue, LNAI 5246,* Springer-Verlag, 2008, p. 277-284, ISSN 0302-9743.

[5] CERVA, P., NOUZA, J. MAP Based Speaker Adaptation in Very Large Vocabulary Speech Recognition of Czech. *Radioengineering*, September 2004, Vol. 13, No 3, p. 42-46, ISSN 1210-2512.

[6] Rajnoha,J., Pollak,P. ASR Systems in Noiosy Enviroment : Analysis and Solutions for Increasing Noise Robustness. *Radioengineering*, 2011, vol. 20, no. 1, p. 74 - 84.

[7] Prochazka,V., Pollak,P. Performance of Czech Speech Recognition with Language Models Created from Public Resources. *Radioengineering*, 2011, vol. 20, no. 4, p. 1002 - 1008.

[8] Leggetter,C.J.,Woodland,P.C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, 1995.

[9] R. M. Schwartz, Y. Chow, S. Roucos, M. Krasner, and J. Makhoul. Improved hidden Markov modeling of phonemes for continuous speech recognition. In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing 1984, pages 35.6.135.6.4, 1984.

[10] Young,S. *The HTK Book for HTK Version 3.4.* Cambridge University Engineering Department, 2006.

[11] Psutka, J., Müller, L., Matoušek, J., Radová, V., *Mluvíme s počítačem česky*, Academia, Praha 2006, p.47 - 56, ISBN:80-200-1309-1.

[12] SPEECON database distributed through the European Language Resources Association [Online]. Available at : http://catalog.elra.info/product_info.php?products_id=1095

[13] CNK-SYN2006PUB, Institute of the Czech National Corpus FF UK, Prague, 2006 [Online]. Available at : http://ucnk.ff.cuni.cz/english/syn2006pub.php,