Improving the Slovak LVCSR Performance by Cluster-Sensitive Acoustic Model Retraining

Peter VISZLAY, Marek ECEGI, Jozef JUHAR

Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Park Komenskeho 13, 041 20 Kosice, Slovak Republic

peter.viszlay@tuke.sk, marek.ecegi@student.tuke.sk, jozef.juhar@tuke.sk

DOI: 10.15598/aeee.v13i4.1448

Abstract. *In this paper, we present a cluster-dependent adaptation approach for HMM-based acoustic models. The proposed approach employs clustering techniques to group the original training utterances into clusters with predefined number. The clustered speech data are intended to adapt an initially pretrained acoustic model to the specific cluster by reestimation based on the standard Baum-Welch procedure. The resulting model, adapted to the homogeneous data may markedly improve the baseline recognition rate, whereas the model complexity may be reduced. In the recognition step, the test samples are scored by each adapted model and the most accurate one is chosen. The proposed approach is thoroughly evaluated in Slovak triphone-based large vocabulary continuous speech recognition (LVCSR) system. The results prove that the cluster-sensitive retraining leads to significant improvements over the baseline reference system trained according to the conventional training procedure.*

Keywords

Acoustic model, adaptation, cluster analysis, reestimation, weighted mean vector.

1. Introduction

An acoustic model (AM) plays an important role in any large vocabulary continuous speech recognition (LVCSR) system because its quality highly affects the overall performance. Several approaches were developed in the past to improve the baseline recognition by AM refinement. One of the most effective and powerful approaches is the AM adaptation. In that case, a general model is adapted to the specific domain (gender, speaker, acoustic conditions, etc.) by advanced methods. Most popular adaptation methods are MLLR (maximum likelihood linear regression), MAP (maximum a posteriori) [1] and eigenvoices [2].

Besides these common adaptation methods, other strategies, such as clustering, are also employed to improve the acoustic model performance. Authors in [3] generated triphone clusters using decision tree based clustering for zero-resourced language of Bengali. The clusters were used to generate tied-state triphones. Other approach to decision tree tying was presented in [4], where the authors employed segmental clustering of acoustic model components in LVCSR system. As was shown in [5], clustering may be applied to compact the acoustic model built from bootstrap to a reasonable size, whereas multiple distance measures for clustering with optimization were investigated. Another approach is focused on retraining, where the parameters of the original model are reestimated with using the adaptation data. This strategy is often used in cross-language modeling tasks for zero-resourced languages, where the existing model of low-resourced language is retrained on the untranscribed audio data [6].

In this paper, we present a fusion of the mentioned cluster analysis and acoustic model retraining without using any typical adaptation method. We utilize clustering to group the training set into crisp clusters, to which is the general acoustic model adapted through the standard Baum-Welch reestimation procedure. We prove that the resulting model may significantly increase the overal performance, whereas the model size and its complexity may be reduced. The LVCSR system evaluation show that the proposed method is effective and it reduces the reference word error rate.

In Section 2. , the clustering is described. Section 3. gives a description of the proposed method. The experimental setup is given in Section 4. The results are presented in Section 5. and finally, the paper is concluded in Section 6.

2. Clustering Approaches

Clustering [7], [8], also known as unsupervised classification is an important problem in pattern recognition field. Clustering partitions the input space into K regions according to some similarity or dissimilarity measure, where the value of K may be known a priori. The aim of clustering is to find a partition matrix $\mathbf{U}(X)$ of the given dataset X, where $X = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ such that $\sum_{j=1}^{n} u_{kj} \geq 1$ for $k = 1, \ldots, K$, $\sum_{k=1}^{K} u_{kj} = 1$ for $j = 1, \ldots, n$ and $\sum_{k=1}^{K}\sum_{j=1}^{n} u_{kj} = n$, where u_{kj} is the membership of pattern $\vec{x}_j$ to cluster C_k. The partition matrix $\mathbf{U}(X)$ of size $K \times n$ may be represented as $\mathbf{U} = [u_{kj}]$, where $k = 1, \ldots, K$ and $j = 1, \ldots, n$. Note that $u_{kj} = 1$ if $\vec{x}_j \in C_k$, otherwise $u_{kj} = 0$ [7].

In this section, we discuss several well-known partitional clustering techniques used in this study. These techniques include K-means clustering, Fuzzy C-means clustering, PAM (partitioning around medoids) and finally, EM (expectation-maximization) model-based clustering.

2.1. K-Means Clustering

The K-means algorithm [8], [9] is an iterative clustering technique that evolves K crisp, compact, and hyperspherical clusters such that the measure

$$J = \sum_{j=1}^{n}\sum_{k=1}^{K} u_{kj} \cdot D^2(\vec{x}_j - \mu_k) \tag{1}$$

is minimized, where

$$\mu_k = \frac{1}{|C_k|}\sum_{i \in C_k} \vec{x}_i \tag{2}$$

is the k-th cluster centroid, $|C_k|$ is the number of points and $\bar{x}_i$ are the points belonging to cluster C_k, respectively. Note that n is the number of all points in the data set. The algorithm may converge to values that are not optimal, depending on the choice of the initial cluster centers. K-means is also not robust to outliers.

2.2. PAM Clustering

PAM clustering, also known as K-medoid clustering [10] is an extension of the K-means algorithm, where medoids are used instead of the cluster means. It tries to minimize the total squared error of the whole data set. It is more robust to noise and outliers as compared to K-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances. The steps of the K-medoid clustering technique closely follow those in K-means.

2.3. Fuzzy C-Means Clustering

Fuzzy C-means clustering [7], [10], [11] is a widely used and powerful unsupervised method that employs the principles of fuzzy sets to find a fuzzy partition matrix. Objects on the boundaries between several clusters are not forced to fully belong to one of the clusters, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. The minimizing criterion used to define good clusters for Fuzzy C-means partitions is defined as:

$$J_\mu(\mathbf{U}, Z) = \sum_{i=1}^{C}\sum_{k=1}^{n}(u_{ik})^\mu D^2(\vec{z}_i, \vec{x}_k), \tag{3}$$

where $\mathbf{U}$ is a fuzzy partition matrix, $\mu \in [1, \infty]$ is the weighting exponent on each fuzzy membership, $Z = [\vec{z}_1, \ldots, \vec{z}_C]$ are C cluster centers and $D(\vec{z}_i, \vec{x}_k)$ is the distance of $\vec{x}_k$ from the i-th cluster center. According to [12], if $D(\vec{z}_i, \vec{x}_k) > 0$ for all i and k, then $(\mathbf{U}, Z)$ may minimize J_μ only if $\mu > 1$ and

$$u_{ik} = \frac{1}{\sum_{j=1}^{C}\left(\frac{D(\vec{z}_i, \vec{x}_k)}{D(\vec{z}_j, \vec{x}_k)}\right)^{\frac{2}{\mu-1}}}, \tag{4}$$

for $1 \leq i \leq C$, $1 \leq k \leq n$ and

$$\vec{z}_i = \frac{\sum_{k=1}^{n}(u_{ik})^\mu \vec{x}_k}{\sum_{k=1}^{n}(u_{ik})^\mu}, \tag{5}$$

where $1 \leq i \leq C$. A common strategy for generating the approximate solutions of the minimization problem in Eq. (3) is by iterating through Eq. (4) and Eq. (5) (also known as the Picard iteration technique) [12].

2.4. EM Clustering

This type of clustering assumes that the clusters follow some specific probability distribution and it is based on mixture models. It aims to determine the parameters of the probability distribution which have the maximum likelihood of their attributes [7]. This algorithm assumes a GMM (Gaussian Mixture Model) with K mixtures and its mixture weights π_k, mean vectors μ_k and covariance matrices Σ_k. Two steps are executed in each iteration; E-step (expectation), where the probability of each point belonging to each cluster is calculated. The second one is the M-step (maximization), which re-estimates the parameter vector of the probability distribution of each class [13]. The cost function of the clustering algorithm is defined as

$$J = \ln\prod_{i=1}^{n} f(x_i) = \sum_{i=1}^{n}\ln\sum_{k=1}^{K}\pi_k f_k(x_i), \tag{6}$$

where $f(x_i)$ is a Gaussian mixture density and $f_k(x_i)$ is the k-th mixture component [14]. The EM clustering assumes the normal distribution of the clusters. If clusters do not follow this distribution, the EM algorithm will fail in providing the appropriate partitioning [7].

2.5. Internal Cluster Validation

The process of evaluating the results of a clustering algorithm is called cluster validity assessment. The so called validation indices are used for measuring the "goodness" of a clustering result comparing to other ones which were created by other clustering algorithms, or by the same algorithms but using different parameter values. In our work, the Dunn index [15] was used to perform the cluster validation:

$$D = \min_{i=1\ldots K}\left\{\min_{j=i+1\ldots K}\left\{\frac{d(C_i,C_j)}{\max\limits_{k=1\ldots K} diam(C_k)}\right\}\right\}, \quad (7)$$

where $d(C_i, C_j)$ is the distance between clusters and $diam(C_k)$ is the maximum cluster diameter.

3. Cluster-Sensitive Acoustic Model Retraining

3.1. Standard LVCSR System

In order to incorporate the clustering-based AM retraining into the standard LVCSR system, we had to modify its baseline components. Therefore, we firstly describe the standard LVCSR system illustrated by Fig. 1. It can be seen that the acoustic front-end is responsible for the appropriate feature extraction and transformation, if it is needed. The features are fed to the decoder, where the most likely hypothesis is find with using the vocabulary and the statistical knowledge from acoustic and language models. The knowledge sources have to be trained beforehand employing well-known training procedures.

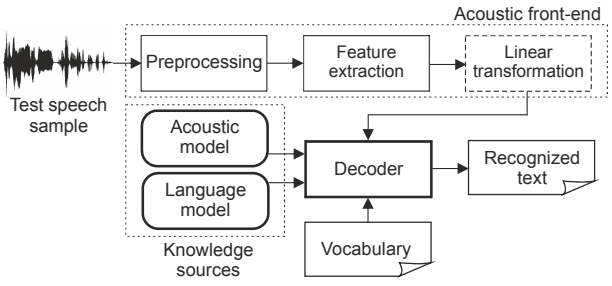

Fig. 1: Block diagram of a general LVCSR system.

3.2. Cluster-Sensitive Training

The aim of the proposed clustering-dependent AM retraining is to partition the complete training set into K disjoint clusters, whereas a cluster contains recordings with similar statistical attributes. The clusters are identified by clustering algorithms described in Section 2. In the next step, an initially trained acoustic model is adapted to the clustered speech data. In this work, each training recording was represented by one-state GMM (Gaussian Mixture Model) described by a probability $b(o_t)$ of generating an observation o_t:

$$b(o_t) = \sum_{m=1}^{M} \pi_m \mathcal{N}(o_t; \mu_m; \Sigma_m), \quad (8)$$

where M is the number of mixture components, π_m is the weight of the m-th mixture and $\mathcal{N}(o_t; \mu; \Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ [16]. Note that we used $M = 16$ mixtures in all GMMs. The parameters were computed by EM algorithm (see Section 2.4.). The GMM computation produced mean and covariance mixture matrices of dimension 16×39 (mixtures $\times$ dimension of MFCCs). In order to perform clustering, it is necessary to find appropriate statistical representatives (vectors) of GMM matrices. Therefore, we suppose to compute weighted mean vector (WMV) of each GMM matrix as [17]:

$$\vec{\mu} = \sum_{m=1}^{M} \pi_m \mu_m, \quad (9)$$

where π_m are weights and μ_m are mixture means. The WMV vectors were then used as input vectors for the subsequent clustering.

It is apparent from the procedure that the most important aspect in our adaptation is the clustering of WMV vectors. We have focused on four different numbers of clusters for each clustering algorithm ($K = 2$, $K = 3$, $K = 5$ and $K = 10$). The determination of the maximum number of clusters was conditioned by value of minimum number of recordings in one cluster and along with the total number of training recordings. We expect that in case of larger number of clusters ($K > 10$), undercounted clusters might be produced and the reestimation can not be done effectively. As was mentioned before, the same clustering algorithm may converge to different cluster configurations at each run because the result is dependent on the initial choice of the parameters. For that reason, the cluster analysis was carried out 10-times and the best one was selected. The selection criterion was based on internal validation with the Dunn index (Eq. (7)). The clustering may also result in incorrect clusters in terms of outlying data elements with very small cluster count. Therefore, we defined the minimum count of each cluster with value $|C_{k_{min}}| = 2500$ recordings (≈ 5 % of the complete set).

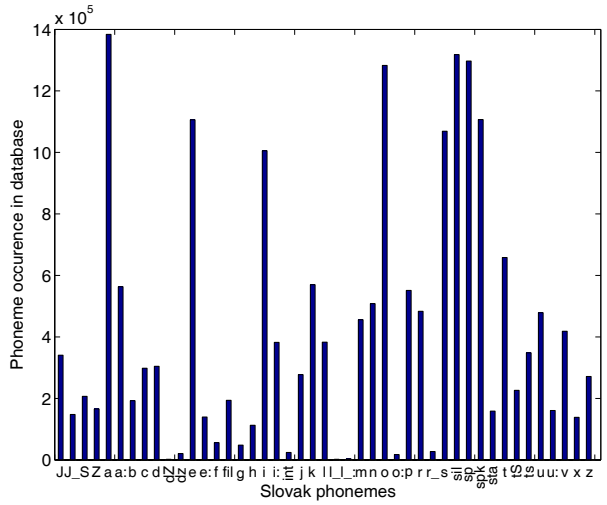

Fig. 2: Occurrences of Slovak phonemes in the training set.

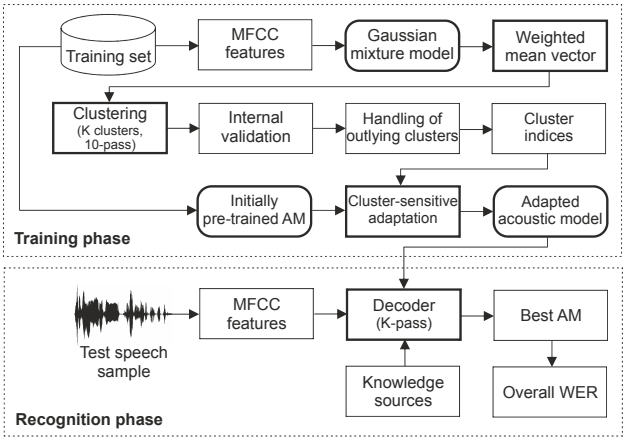

Fig. 3: Block diagram of training and recognition phase of the proposed adaptation approach.

The outlying clusters were joined to the nearest correct one in terms of minimum Euclidean distance.

Regarding the phonetic balance of the resulting clusters, we carried out an extensive phonetic analysis of the whole training part (see Fig. 2). This chart describes real statistical counts of Slovak phonemes, including the noise-specific phones [18]. It is obvious that the training data are not phonetically balanced because they represent real attributes of the Slovak language. Note that the data were not manually balanced afterwards. There is a high degree of variability between the counts caused by the occurrence in the real speech. The highest counts (more than 1 million occurrences) are typical for vowels and noise phones and lower counts (around 300 000 occurrences and less) are typical for consonants. If we consider this nature of training data and if we further consider the fact that each cluster contains a sufficient amount of data ($|C_{k_{min}}| = 2500$), we expect that the correct clusters follow the same or very similar phonetic distribution, probably with slight count variations (depending on K). In other words, we assume that the clusters are not phonetically balanced.

It is hard to determine how the LVCSR performance is affected by the phonetic distribution in each cluster. In order to determine the influence, a comprehensive performance analysis would be required. We assume that the phonetic balance of the cluster does not affect the overall performance markedly, while reasonable phoneme counts in each cluster are kept.

The correct clusters are finally intended to acoustic model adaptation. It should be clarified that the parameters of the original AM (probabilities and mixtures of HMM) are adjusted and reestimated with using the adaptation data of the specific speech cluster. We employed the standard Baum-Welch reestimation procedure to compute the new parameters [16]. To sum up, we do not utilize any typical adaptation algorithm (MLLR, MAP, EV) to adapt an acoustic model.

We have also focused on the effect of the quality of the initial AM to the overall adaptation performance. In the most common adaptation tasks, a general AM, trained on the complete training set is usually adapted to the desired domain. In our case, the general AM is denoted as the reference AM. However, we found that the adaptation of a weak initial AM, just pre-trained on randomly selected training subset (e.g. 50 % of the complete set) holds the key of considerably increased LVCSR performance. This interesting fact also introduces some benefits of our adaptation approach, e.g. an adapted AM, originally pre-trained just from 25 % data, may achieve markedly lower WER than the AM, originally trained on the full data. In our evaluation, we have analysed four partially pre-trained AMs: $P = 10$ %, $P = 25$ %, $P = 50$ % and $P = 75$ %, where P defines the size of the randomly selected subset.

3.3. Modified Recognition Phase

In order to evaluate the proposed method, it was necessary to modify the standard recognition process. Compared to the standard LVCSR system, the modification was focused on the decoding because it is required to perform $K-$pass decoding for each test sample, where K is the number of adapted AMs. In each pass, the word level error rate is computed using the reference transcriptions and after all passes, the minimum level of WER is determined and accumulated. This procedure is repeated for each recording. Finally, the overall LVCSR performance is evaluated in form of global WER computed by averaging of the accumulated WER levels. The training and recognition phase of our adaptation approach is depicted in Fig. 3 in detail.

At the end, it is interesting to compare the standard LVCSR system with the modified one, based on AM

retraining. It can be seen that the training phase of the standard system (Fig. 1) is extended by clustering-related steps and AM retraining (Fig. 3). As we mentioned, the recognition requires K-pass decoding with separate WER evaluation in each pass, whereas the best adapted AM is chosen for each test recording. This is the main reason, why the proposed method performs better than the standard one.

4. Experimental Setup

The Slovak parliamentary corpus *ParDat1* [19] used in our study contains approx. 100 hours of spontaneous parliamentary speech. The training part involves 50876 utterances collected from 120 speakers ($\approx$ 90 % of men). The testing database includes another 884 phonetically balanced recordings with total duration up to 3 hours.

Throughout the experiments, the standard MFCC (Mel Frequency Cepstral Coefficient) features with cepstral mean normalization (CMN) were extracted, including their first and second derivatives and log energy, resulting in a 39-dimensional vectors.

The LVCSR system employed cross-word, three-state, left-to-right structure tied-state context-dependent triphone HMM (Hidden Markov Model) acoustic models. All acoustic models were trained in the maximum likelihood (ML) sense with GMM (Gaussian Mixture Model) density functions. At the end of the ML training process, about 12000 final triphones were produced and modeled with 32 Gaussians per state for each acoustic model, according to the reference training setup of the HTK toolkit [16]. The LVCSR decoder employed a bigram language model [20] and vocabulary containing approximately 125000 unique, phonetically transcribed words.

For LVCSR system evaluation, we chose the word-level error rate (WER) computed as:

$$\text{WER } [\%] = \frac{S + D + I}{N} \cdot 100, \tag{10}$$

where S represents the substitutions, D is the number of deletions, I is the number of insertions and K is the total number of reference words [16].

Finally, we note that the computing of weighted mean vectors, clustering, internal cluster validation, handling of outlying clusters and the evaluation were carried out in the Matlab programming environment. On the other hand, the feature extraction, GMM modeling, acoustic model training and retraining and the decoding were performed using the HTK Toolkit.

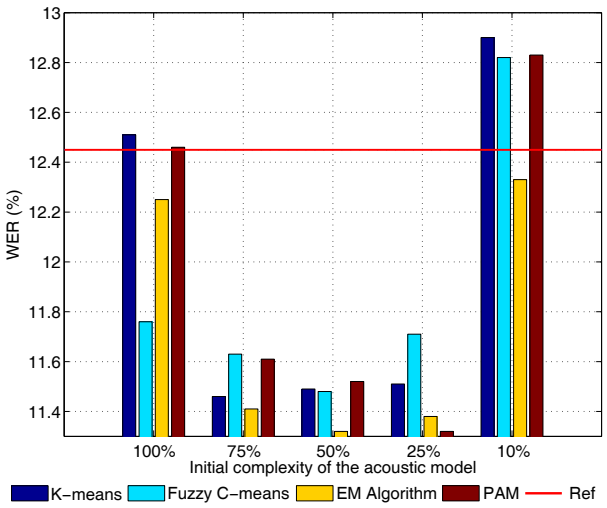

Fig. 4: WER levels for adapted LVCSR systems, 2 clusters.

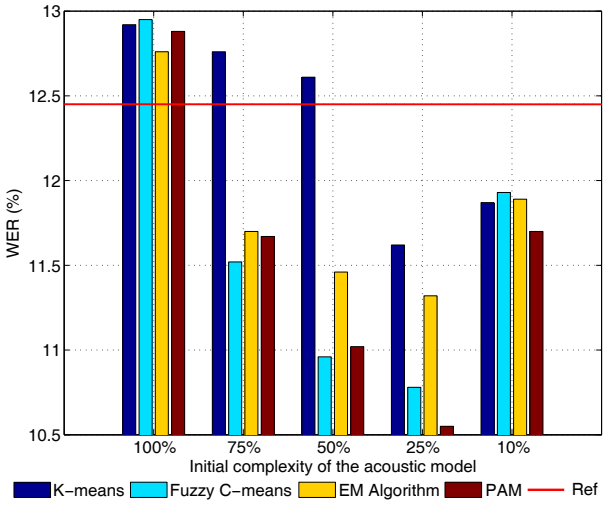

Fig. 5: WER levels for adapted LVCSR systems, 3 clusters.

5. Experimental Evaluation

The experimental results are given in Fig. 4, Fig. 5, Fig. 6, Fig. 7. The performance of the reference LVCSR system, trained on standard MFCCs, is depicted with red line and its value is $\text{WER}_{\text{ref}} = 12.45$ %. Thus, each value of WER falling below the red line means satisfactory result. The reference acoustic model was trained from the complete set ($P = 100$ %). At first, if we compare the results for $K = 2$ clusters in Fig. 4, we can observe that the reference WER is decreased for all clustering methods at the same time only if $P =$ 25 % up to 75 %. In other cases, the reference WER is improved only for Fuzzy C-means and EM clustering. The minimum value of WER for this setup is 11.32 % for EM clustering, thus the WER_{ref} was reduced by -1.13 %.

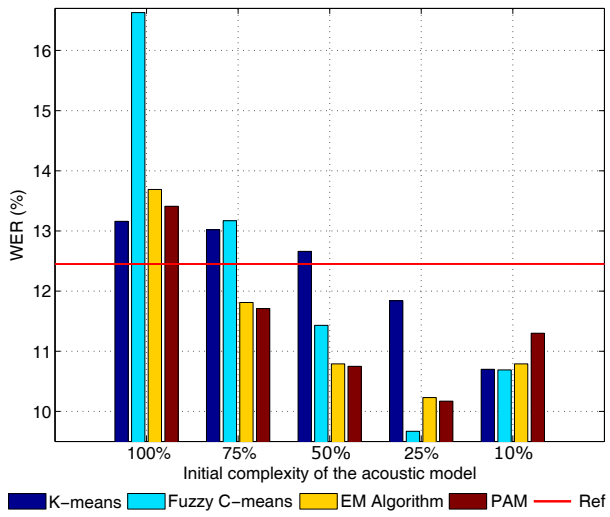

Fig. 6: WER levels for adapted LVCSR systems, 5 clusters.

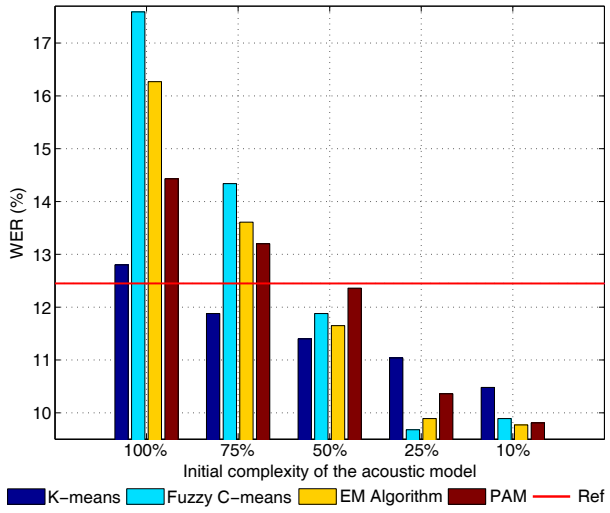

Fig. 7: WER levels for adapted LVCSR systems, 10 clusters.

In case of adaptation to $K = 3$ clusters (Fig. 5), the reference WER is reduced for all methods for initial models with $P = 25$ % and 10 %. The minimum value of WER for this setup is 10.55 % for PAM, thus the $\mathrm{WER}_{\mathrm{ref}}$ was reduced by -1.90 %.

The adaptation to 5 clusters (Fig. 6) has very similar nature. The highest reduction in WER was measured for Fuzzy C-means and initial model $P = 25$ % and its value is 9.67 %. This value concurrently represents the absolute minimum value of WER achieved by the proposed method in the whole evaluation. In that case, the value of $\mathrm{WER}_{\mathrm{ref}}$ was reduced exactly by -2.78 %. This means a relative LVCSR performance improvement by 22.33 %.

Finally, from the chart in Fig. 7 it is evident that the adaptation to 10 clusters clearly outperformed the reference system for all methods in case of $P = 10$ %, 25 % and 50 %. For greater values of P, the values of WER began to rise. This adaptation yielded minimum value of WER 9.68 % for Fuzzy C-means and initial model $P = 25$ % again ($\mathrm{WER}_{\mathrm{ref}}$ improved by -2.77 %).

From a global point of view we can conclude that the lowest values of WER were achieved through EM and Fuzzy C-means clustering and most often by initial AMs with $P = 25$ %. We state that this type of AM is the most suitable for cluster-sensitive adaptation. We can also observe that initial AMs with $P = 10$ % and $P = 50$ % yield partially great improvement, too. We found that the number of clusters has not a crucial impact to the overall performance. It seems that the optimal values of K are $K = 5$ and $K = 10$. Note that the initial AM, trained on the complete set ($P = 100$ %) gives after adaptation the worst results almost in all experiments, without respect to the clustering. We have proven that for our adaptation approach it is sufficient to use a weak, non-precise AM, which yields significantly lower levels of WER than the fully-trained adapted AM. Additionally, the adaptation of less complex initial AM is also less computationally expensive, which is a much desired feature for LVCSR systems.

In order to declare the effectiveness of the presented adaptation approach, we contrast the performance of our adapted LVCSR system with two related, recently published works, where similar LVCSR systems employing conventional adaptation techniques were described. The first work [21] is focused on MLLR-based speaker adaptation task for Czech LVCSR system with two different clustering methods (knowledge-based and automated one). The authors declare here relative improvements in the range of 16.68 % up to 20.91 %, depending on the clustering method and the number of regression trees for MLLR. The second one [22] presents an on-line adaptation using KSVD-based acoustic clustering for real-time applications, where the adaptation performance in UK English LVCSR task was evaluated. The authors reported that the adaptation approach is capable of providing a 6 % relative WER reduction, rather in range of 2.0 % up to 6.1 %, whereas WER increasing was also observed. It can be concluded that the performance of the presented adaptation approach based on model retraining is competitive with other state-of-the-art adaptation techniques.

6. Conclusion and Future Work

In this work, we presented a cluster-sensitive adaptation for HMM-based acoustic models. We proved that our adaptation is able to reduce the reference WER significantly. This fact suggests the suitability of this method for LVCSR systems. We intend further to re-

fine the recognition process by selection the appropriate AM without necessity to perform K-pass decoding.

Acknowledgment

The research presented in this paper was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the project VEGA1/0075/15 (50 %) and Research and Development Operational Program funded by ERDF under the project UVP Technicom ITMS-26220220182 (50 %).

References

[1] WANG, Z., T. SCHULTZ and A. WAIBEL. Comparison of acoustic model adaptation techniques on non-native speech. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*. Hong Kong: IEEE, 2003, pp. 540–543. ISBN 0-7803-7663-3. DOI: 10.1109/ICASSP.2003.1198837.

[2] KUHN, R., P. NGUYEN, J. JUNQUA, L. GOLDWASSER, N. NIEDZIELSKI, S. FINCKE, K. FIELD and M. CONTOLINI. Eigenvoices for speaker adaptation. In: *Proceedings of the International Conference on Spoken Language Processing*. Sydney: Australian Speech Science and Technology Association, 1998, pp. 1–4. ISBN 1-876-346-175.

[3] BANERJEE, P., G. GARG, P. MITRA and A. BASU. Application of triphone clustering in acoustic modeling for continuous speech recognition in Bengali. In: *Proceedings of the International Conference on Pattern Recognition*. Tampa: IEEE, 2008, pp. 1–4. ISBN 978-1-4244-2174-9. DOI: 10.1109/ICPR.2008.4761657.

[4] REICHL, W. and W. CHOU. Decision tree state tying based on segmental clustering for acoustic modeling. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*. Seattle: IEEE, 1998, pp. 801–804. ISBN 0-7803-4428-6. DOI: 10.1109/ICASSP.1998.675386.

[5] CHEN, X., C. XIAODONG, J. XUE, P. OLSEN, J. HERSHEY, B. ZHOU and Y. ZHAO. Clustering of bootstrapped acoustic model with full covariance. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Prague: IEEE, 2011, pp. 4496–4499. ISBN 978-1-4577-0538-0. DOI: 10.1109/ICASSP.2011.5947353.

[6] LOOF, J., C. GOLLAN and H. NEY. Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a polish speech recognition system. In: *Proceedings of InterSPEECH*. Brogthon: IEEE, 2009, pp. 88–91. ISBN 978-1-61567-692-7.

[7] BANDYOPADHYAY, S. and S. SAHA. *Unsupervised classification*. Berlin Heidelberg: Springer-Verlag, 2013. ISBN 978-3-642-32450-5.

[8] JAIN, A. K. and R. C. DUBES. *Algorithms for clustering data*. Upper Saddle River: Prentice-Hall, 1988. ISBN 978-0130222787.

[9] EVERITT, B. S., S. LANDAU and M. LEESE. *Cluster Analysis*. London: Wiley, 2001. ISBN 978-0470749913.

[10] THEODORIDIS, S. and K. KOUTROUMBAS. *Pattern Recognition*. 3rd ed. Orlando: Academic Press, 2006. ISBN 978-0-12-369531-4.

[11] SUGANYA, R. and R. SHANTHI. Fuzzy C-Means Algorithm - A Review. *International Journal of Scientific and Research Publications*. 2012, vol. 2, iss. 11, pp. 1–3. ISSN 2250-3153.

[12] BEZDEK, J. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Springer, 1981. ISBN 978-1-4757-0452-5.

[13] DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*. 1977, vol. 39, no. 1, pp. 1–38. ISSN 1467-9868.

[14] REYNOLDS, D. A. and R. C. ROSE. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*. 1995, vol. 3, iss. 1, pp. 72–83. ISSN 1063-6676. DOI: 10.1109/89.365379.

[15] LEGANY, C., S. JUHASZ and A. BABOS. Cluster validity measurement techniques. In: *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*. Wisconsin: World Scientific and Engineering Academy and Society, 2006, pp. 388–393. ISBN 111-2222-33-9.

[16] YOUNG, S., G. EVERMANN, M. J. F. GALES, T. HAIN, D. KERSHAW, X. LIU, G. MOORE, J. ODELL, D. OLLASON, D. POVEY, V. VALTCHEV and P. C. WOODLAND. *The HTK Book (for HTK Version 3.4)*. Cambridge: Cambridge University Engineering Department, 2006.

[17] DE LEON, P. L. and V. APSINGEKAR. Reducing speaker model search space in speaker identification. In: *Proceedings of Biometrics Symposium*. Baltimore: IEEE, 2007, pp. 1–6. ISBN 978-1-4244-1549-6. DOI: 10.1109/BCC.2007.4430544.

[18] RUSKO, M. INCO-COPERNICUS-977017-ED1.12.3. *Definition of Corpus, scripts, standards and Specifications of environmental and speaker specific coverage applied to the Slovak speech database*. Bratislava: Slovak Academy of Sciences, Slovakia, 1999.

[19] DARJAA, S., M. CERNAK, M. TRNKA, M. RUSKO and R. SABO. Effective triphone mapping for acoustic modeling in speech recognition. In: *Proceedings of INTERSPEECH*. Florence: International Speech Communication Association, 2011, pp. 1717–1720. ISBN 978-1-61839-270-1.

[20] STAS J., D. HLADEK and J. JUHAR. Recent advances in the statistical modeling of the slovak language. In: *Proceedings of the 56th International Symposium ELMAR'14*. Zadar: IEEE, 2014, pp. 39–42. ISBN 978-953-184-199-3. DOI: 10.1109/ELMAR.2014.6923310.

[21] BORSKY, M. and P. POLLAK. Knowledge-based and automated clustering in MLLR adaptation of acoustic models for LVCSR. In: *Proceedings of the International Conference on Applied Electronics*. Pilsen: IEEE, 2012, pp. 33–36. ISBN 978-1-4673-1963-8.

[22] SHAHNAWAZUDDIN, S. and R. SINHA. Fast on-line adaptation using KSVD based acoustic clustering. In: *Annual IEEE India Conference INDICON'13*. Mumbai: IEEE, 2013, pp. 1–5. ISBN 978-1-4799-2274-1. DOI: 10.1109/INDCON.2013.6725938.

About Authors

Peter VISZLAY was born in Rozvnava, Slovak Republic in 1985. He received his M.Sc. in Electronics and Telecommunication from the Technical University of Kosice in 2009. He received Ph.D. degree in Infoelectronics from the same university in 2013. Currently, he works as a researcher at the Department of Electronics and Multimedia Communications. He is involved in speech processing, linear feature transformations, acoustic modeling, continuous speech recognition systems and speech separation methods.

Marek ECEGI was born in Vranov nad Toplou, Slovak Republic in 1991. He received his B.Sc. and M.Sc. in Telecommunications in 2013 and 2015 from the Technical University of Kosice, respectively. He is a Ph.D. student at the Deptartment of Electronics and Multimedia Communications. He is interested in speech processing and clustering methods.

Jozef JUHAR was born in Poproc, Slovakia in 1956. He graduated from the Technical University of Kosice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Kosice in 1991. Now he works as a full professor and as a Head of Department at the Department of Electronics and Multimedia Communications. He is author and co-author of more than 200 scientific papers. His research interests include digital speech and audio processing, speech/speaker identification, speech synthesis, development in spoken dialogue and speech recognition systems in telecommunication networks.