Multi-Conditionally Trained ASR System for Reverberant Speech Captured by Spherical Microphone Array in Adverse Acoustic Conditions

Peter Viszlay, Ján Staš, Martin Lojka, Jozef Greššák, Jozef Juhár, Slavomír Gereg

Department of Electronics and Multimedia Communications,
Technical University of Košice, Park Komenského 13, 042 00 Košice, Slovakia
{peter.viszlay, jan.stas, martin.lojka, jozef.gressak, jozef.juhar}@tuke.sk, slavomir.gereg@student.tuke.sk

Abstract

This paper addresses the complex problem of speech dereverberation in terms of feature enhancement, employed in automatic speech recognition system for adverse acoustic conditions. The presented system uses a spherical microphone array to capture the meeting speech in a medium sized, acoustically not optimized room. The reverberant speech is enhanced in the mel-spectral domain by using a distribution matching of long time context spectro-temporal supervectors of speech that are decorrelated by principal component analysis. The enhanced features are further improved by our originally proposed class-dependent two-dimensional linear discriminant analysis. To minimize the mismatch between the training and testing conditions, we performed multi-condition training with artificially reverberated data. The proposed system is evaluated on Slovak large vocabulary continuous speech recognition task at different stages. Although the results show absolute word error rate decrease by 25%, there is still a room for system tuning and improvements in this challenging task.

1. Introduction

Reverberation robust automatic speech recognition (ASR) has been a challenging issue in recent decades for many researchers and still remains a subject of major interest. It is known that ASR systems are quite sensitive to reverberant conditions that cause performance degradation (Wölfel and McDonough, 2009), (Mitra et al., 2015). The primary reason is the acoustic channel mismatch between training and testing conditions (Kalinli et al., 2010), (Dennis and Dat, 2015). Many methods have been proposed in the literature to circumvent reverberation effects and improve the robustness of ASR. They can be grouped under two main categories: feature enhancement algorithms applied before the recognition of the corrupted speech and approaches to minimize the acoustic-condition mismatch (Kalinli et al., 2010), (Yoshioka et al., 2012). They can also be combined together with more or less advantageous results (Rajnoha, 2009). A subcategory is represented by cases when the speech is captured by a microphone array, where the negative reverberation effects in adverse conditions can initially be suppressed by advanced microphone-array processing (Brandstein and Ward, 2001).

The mentioned feature enhancement techniques can operate in the spectrum or in the feature domain directly. A class of algorithms can use a prior speech model that assists in the enhancement process (Kalinli et al., 2010). On the other hand, multi-condition techniques are successfully employed to reduce the acoustic-condition mismatch by interfering the clean training data with room impulse responses (RIRs) (Mitra et al., 2014), (Ribas et al., 2015).

In this work, we were motivated by the REVERB challenge that was primarily oriented on the reverberation robust ASR (Kinoshita et al., 2013). This contribution proposes a system that combines speech dereverberation for spherical microphone array with the multi-condition training strategy and aims to reduce the word error rate (WER) as much as possible. The dereverberation method performs feature enhancement through distribution matching of spectro-temporal supervectors that are decorrelated by principal component analysis (PCA). It was proposed by (Palomäki and Kallasjoki, 2014) within the REVERB challenge and slightly adopted by us. The proposed system is evaluated in Slovak large vocabulary continuous speech recognition (LVCSR) for meeting speech captured in a medium sized meeting room. We have built the system upon our previous works that introduced approaches for microphone array processing (Hiľovský et al., 2016), discriminative feature transformation (Viszlay et al., 2014), decoding (Staš et al., 2015) and language model adaptation (Staš et al., 2017). Although the achieved results show absolute WER decrease by 25%, we are aware that there is still a room for system tuning and improvements.

In Section 2., the microphone array processing is described. Section 3. describes the dereverberation method. Section 4. reviews linear transformations and Section 5. presents the multi-condition framework. The experimental setup is given in Section 6. and the evaluation is presented in Section 7. Finally, the paper is concluded in Section 8.

2. Microphone array signal processing

The presented ASR system uses a 32-channel spherical microphone array EM32 Eigenmike[1] with sphere diameter of 8.4 cm. To control the recorded multichannel signal, we used our software library MAPL (Microphone Array Processing Library) (Hiľovský et al., 2016) that implements algorithms for localization of sources (Nadiri and Rafaely, 2014) and beamforming (Rafaely, 2015).

The localization block is based on time-frequency analysis and direct path dominance test computed in four stages: spatial transformation into spherical harmonic domain using short-time Fourier transform; spatial correlation; direct path dominance test and the estimation of the direction of arrival (DOA). The information from the localization block is used in the beamformer to adapt the directional characteristics for each sound source.

[1] https://mhacoustics.com/products

3. Speech dereverberation

The dereverberation method used in this work is an unsupervised approach that utilizes clean speech prior. The estimated features of the dereverberated speech are obtained by matching the reverberant speech distribution to the clean speech distribution. The distribution matching (DM) aims at recovering the clean spectra $\mathbf{x}$ from the reverberant spectra $\mathbf{y}$, when the clean prior distribution $p(\mathbf{x})$ is known a priori and the observed reverberant speech $p(\mathbf{y})$ can be estimated in the recognition phase (Keronen et al., 2015). The dereverberation can be considered as a Bayesian inverse problem, where the posterior distribution for the dereverberated speech $p(\mathbf{x}|\mathbf{y})$ is defined as:

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x})p(\mathbf{y}|\mathbf{x}), \tag{1}$$

where $p(\mathbf{y}|\mathbf{x})$ represents the reverberant observation (Palomäki and Kallasjoki, 2014). The DM is performed in three steps in two iterations. In the first iteration, the reverberated speech $\mathbf{x}$ is treated as the observed speech $\mathbf{y}$, whereas in the second iteration, the dereverberated estimate is used as the observation $\mathbf{y} = \hat{\mathbf{x}}$.

Firstly, the speech signal is parametrized into mel-spectral feature vectors that are used as the input to the feature enhancement. In order to counteract the long lasting effects of reverberation, the feature vectors are stacked over N consecutive frames to form Nd-dim. temporal context supervectors $\mathbf{y} = [\mathbf{y}_t^T \dots \mathbf{y}_{t+N-1}^T]^T$, where N is chosen with respect to the room impulse responses (RIRs) and d is the number of mel-filters. The speech features $\mathbf{y}$ can be expressed as $\mathbf{y} \approx \mathbf{H}\mathbf{x}$, where $\mathbf{H}$ is a filter matrix performing the convolution on $\mathbf{x}$ constructed from clean features.

Since the constructed supervectors $\mathbf{x}$ and $\mathbf{y}$ are highly correlated along dimensions, PCA is applied to decorrelate the spectro-temporal supervectors on a log-scale as:

$$\mathbf{g}'_{\mathbf{y}} = \mathbf{D}\log\mathbf{y} \approx \mathbf{D}\log\mathbf{H}\mathbf{x}, \tag{2}$$

where $\mathbf{g}'_{\mathbf{y}}$ is the observed supervector in the decorrelated space. The goal of the DM is to develop one-dimensional element-wise bijective mapping function $F_{yx}^{(m)}$ that maps the elements $g'_y(m)$ from the reverberant PCA domain to dereverberated ones $\tilde{g}'_x(m)$ as:

$$\tilde{g}'_x(m) = F_{yx}^{(m)}\big(g'_y(m)\big), \tag{3}$$

where m is the element index. The mappings can be obtained if the distributions are represented by inverse cumulative distribution functions. For simplicity, by dropping the indices m, the dereverberated log-spectral supervector $\tilde{\mathbf{x}}'$ can be estimated as:

$$\tilde{\mathbf{x}}' = \mathbf{D}^{-1}F_{yx}\big(\mathbf{g}'_y\big), \tag{4}$$

where $\mathbf{D}^{-1}$ is the inverse PCA transformation. Finally, the supervectors $\tilde{\mathbf{x}}'$ are unstacked back to mel-spectral vectors $\tilde{\mathbf{x}}$ that are post-processed by a Wiener filter and used as the input for ASR. The dereverberation process based on the DM method is illustrated In Fig. 1. A detailed mathematical description of the DM method can be found in the original works (Palomäki and Kallasjoki, 2014) and (Keronen et al., 2015). The dereverberation toolkit is publicly available[2] in form of Matlab functions.

[2]http://users.spa.aalto.fi/kpalomak/DM.html

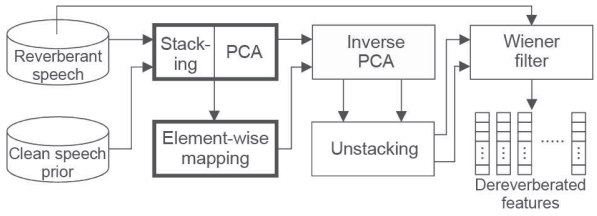

Figure 1: Block diagram of the dereverberation process

4. Extended front-end processing

The discrete cosine transform (DCT) decorrelates the features in the MFCC (mel-frequency cepstral coefficients) analysis within the current frame, while the adjacent frames can be still correlated. We employed PCA (Jolliffe, 1986) to decorrelate all features globally. We computed the global covariance matrix C as:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \tag{5}$$

where n is the number of training vectors, $\mathbf{x}_i$ is the current vector and $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ is the global mean. The obtained the PCA matrix was used to transform the enhanced features instead of the frame-level DCT matrix.

Principal component analysis is a good choice for optimal decorrelation of the features but it may not provide satisfying discriminability of the transformed features because it does not take into account the class membership of the feature vectors. Therefore, we applied class-dependent two-dimensional LDA (CD-2DLDA) to the PCA-based features. We proposed CD-2DLDA in (Viszlay et al., 2014) as an extension of the classical 2DLDA (Ye et al., 2005) on TIMIT (Garofolo et al., 1993) speech recognition task. In this work, we adopted the method to large vocabulary task for Slovak language.

CD-2DLDA employs two-pass recognition strategy, where the first pass generates class labels of test samples using the baseline system that are used to class-dependent transformation. In the second pass, recognition of test samples is performed using CD-2DLDA based acoustic model. The key idea behind CD-2DLDA is to run the original 2DLDA algorithm for each triphone class Π_i separately to obtain transformation matrices L_i and R_i. We defined the class-dependent within-class scatter matrix $S_{w_i}^R$ as:

$$S_{w_i}^R = \sum_{X\in\Pi_i}(X - M_i)RR^T(X - M_i)^T \tag{6}$$

and the class-dependent within-class scatter matrix $S_{w_i}^L$ of class i coupled with L as:

$$S_{w_i}^L = \sum_{X\in\Pi_i}(X - M_i)^T LL^T(X - M_i), \tag{7}$$

where X is a feature matrix, n_i is the number of feature matrices in class i, $M_i = \frac{1}{n_i}\sum_{X\in\Pi_i}X$ is the class mean matrix, and k is the number of classes. and n is the total number of training elements. The between-class scatter matrices S_b^L and S_b^R stay same, as in 2DLDA. The detailed description of CD-2DLDA can be found in the original paper (Viszlay et al., 2014).

5. Multi-condition training strategy

Generally, reverberation causes acoustic mismatch between the training and testing conditions, which usually degrades the ASR performance. ASR systems trained purely on clean data usually show dramatic performance degradation under real conditions (Mitra et al., 2015), (Harper, 2015). Multi-condition training (MCT) is a well-established technique to handle reverberant conditions by augmenting the clean training data with additional degraded data (using RIRs) and by training a reverberant-robust acoustic model that reduces the acoustic-condition mismatch. Moreover, the combination of MCT with front-end speech enhancement generally improves the ASR performance even more (Ribas et al., 2015), (Harper, 2015). Based on the mentioned facts, we have built an ASR system that follows the way of multi-conditional framework. The intention of MCT is to add some reverberation and noise to make the training data "dirty". In this work, we created the MCT data in two different datasets: 1) artificially reverberated dataset and 2) pseudo-clean dataset.

5.1. Artificially reverberated dataset

The first dataset was generated through a convolutive interference of clean speech signals with a FIR filter represented by RIRs. For that purpose, we utilized the Gardner's reverberation toolkit[3] that supports corrupting of speech with a reverberation of small, medium and large room. Several datasets were generated according to different configurations (RIRs). In order to analyse the influence of the different reverberation levels to the error rate, the datasets were explored independently, always joined with the clean dataset. Based on a comprehensive experimental analysis, we chose the best-matching dataset that provided the lowest WER. In other words, we minimized the mismatch between the training and testing conditions.

However, we found that even better channel condition match can be achieved if the reverberated dataset is subsequently processed by the dereverberation algorithm. It is an expectable outcome because the test data were processed by the same algorithm so the acoustic channel conditions can be perfectly matched.

5.2. Pseudo-clean dataset

The second dataset was generated in similar but much more simpler way. Based on supplementary experimental analysis we found that an additional ASR performance gain can be achieved when also the clean speech dataset is processed by the dereverberation algorithm and joined to the multi-conditional dataset. We called that particular dataset as *pseudo-clean* one. In that way, the amount and the level of the MCT training data is increased.

The block diagram of the proposed system is illustrated in Fig 2. The artificial reverberation, feature enhancement for training data, PCA, CD-2DLDA and AM training are performed offline.

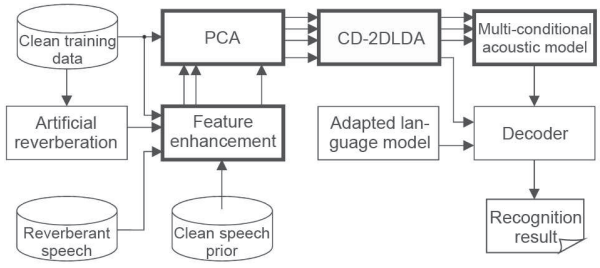

Figure 2: Block diagram of the proposed system

(a) (b)

Figure 3: Spherical microphone array *EM32 Eigenmike* (a) and the recording setup in the meeting room (b)

6. Experimental setup

6.1. Multichannel recording setup

The recording scenario was realized using a spherical microphone array EM32 Eigenmike (see Fig. 3 (a)) in a medium sized meeting room with a relatively long room impulse response. In order to ensure real acoustic environment, the room was not acoustically optimized in any way. The microphone array was set on a fixed stand at a height of 1.75 meters from the floor and placed in the middle of the room (see Fig. 3 (b)). Synchronously with EM32, the speech was recorded also with a close talk microphone AKG C111 LP that was worn by each speaker. This signal was considered as an auxiliary clean speech intended for comparison with the reverberated one. During recording, the speakers sat at the tables around the microphone array and they read the prepared text.

6.2. Acoustic corpus and speech data

We used a gender-balanced acoustic database that included 80 hours of manually annotated speech acquired from TV shows called as "Court Room". It covers overall 98 different speakers.

The test data were generated from the acoustic material recorded by EM32 and processed into single-channel recordings using MAPL. The resulting test set was represented by 329 speech segments that corresponded to 21 minutes of speech. The speaker inventory consisted of 3 speakers (2 males, 1 female) that were not included in the training part. Only one speaker was talking at a time. A selected part of the test set (30 segments) is freely available[4] for demonstration purposes.

[3] www.cps.unizar.es/~fbeltran/matlab_files.html

[4] http://nlp.web.tuke.sk/pages/reverberant

The single-channel source speech signals were pre-emphasized and windowed every 10 ms using Hamming window of length 25 ms. Fast Fourier transform and mel filter-bank analysis with 26 channels were applied to the windowed segments. The subsequent processing steps were performed according to the specific system (FBANK e.g. log-mel-spectral, MFCC, PCA, CD-2DLDA).

6.3. Acoustic modeling

The presented Slovak LVCSR system makes use of triphone context-dependent acoustic models based on three-states left-to-right hidden Markov models (HMMs) with 32 Gaussian mixtures per state. The typical tree-based state tying algorithm for HMMs (Young et al., 2006) has been replaced by an effective triphone mapping (Darjaa et al., 2011) that produced 3177 triphone classes used also in the CD-2DLDA based computing.

6.4. Language modeling

The background language model was created using the SRILM Toolkit (Stolcke, 2002). It was restricted to the vocabulary size of about 500 thousand unique words and smoothed by the Witten-Bell back-off algorithm. The trigram model was trained on the web-based corpus of Slovak written texts. The corpus size of about 1.89 billion tokens and 110.75 million sentences was then segmented into 5.93 million paragraphs with approximately 315 words on average for better representation in the vector space. After that, semantic indexing and vector space modeling were implemented to retrieve a subset of text documents from the background corpus relevant to the topic and speaking style of a speaker. Also, the authors proposed document retrieval approach based on using Paragraph Vectors (Le and Mikolov, 2014) for topic-specific modeling to improve speech recognition accuracy for individual speakers (Staš et al., 2017). In this approach, they select a subset of text documents semantically similar to the output hypotheses from recognized speech segments in the first decoding stage. A small topic-specific LM was then created from the relevant documents, interpolated with the background LM, adapted to the current topic and speaking style of a speaker, and applied during the second decoding stage(s).

6.5. Decoding and evaluation

The speech recognition decoder was based on large vocabulary recognition engine Julius (Lee et al., 2001) that was modified to support multi-threaded parallel speech recognition and sharing acoustic and language models among all instances for memory space saving purposes (Lojka et al., 2014). The acoustic model training and the decoding were run in parallel mode on a high-performance computing cluster IBM Blade System x HS22.

Word error rate (WER) was used to evaluate the performance of the LVCSR system. It was computed by comparing reference annotations against the recognized result as $WER = \frac{S+D+I}{N} \times 100$ [%], where S refers to the number of substituted words, D is related to words, which are missed out, I indicates the number of words incorrectly added by the recognizer, and N is the total number of words in the reference (Young et al., 2006).

Baseline AM (39-dim.)	Eigenmike	CTM
MFCC_0_D_A_Z	83.03	26.63
MFCC_D_A_Z	85.71	32.83

Table 1: Word error rates (%) of the baseline systems

Speech prior	#Recordings	Duration	WER (%)
Court Room	#500	1.19 h	75.98
Court Room	#2000	1.99 h	73.65
Judicial	#1713	2.90 h	73.16
CISCO lect.	#1728	3.31 h	**72.50**
BN-studio	#2115	2.83 h	79.02
BN-exterior	#2577	2.56 h	76.19

Table 2: Performance of the feature enhancement method

7. Experimental results and evaluation

The experiments were performed on Slovak LVCSR task for meeting speech. There are overall five evaluation stages. The results are expressed in terms of WER in %.

7.1. Baseline system

We present our results for two baseline systems, that are listed in Tab. 1. Both systems use standard 39-dim. MFCC coeffs. with a subtle difference in features. The system denoted as MFCC_0_D_A_Z uses 12 cepstral coeffs., 0-th coeff., Δ and Δ^2 coeffs., whereas the model denoted as MFCC_D_A_Z uses 13 cepstral coeffs., Δ and Δ^2 coeffs. and it does not contain the 0-th coeff. It was trained for secondary comparison with the results obtained from feature enhancement (see Section 7.2.). The WERs are really high because the LVCSR system trained on clean data could not recognize the reverberant speech correctly. The MFCC_0_D_A_Z system is treated as the reference in the following sections and would be treated as final one if any improvements were further done. Just for interest, we also show additional results achieved by recognizing the same speech session from the close talk microphone (CTM). It can be clearly seen that the speech almost without reverberation produces significantly lower error rates.

7.2. Feature enhancement

The second stage was the complex evaluation of the dereverberation algorithm on the test data described in Section 6.2. The algorithm generated 26-dim. enhanced features in the log-mel-spectral domain that could not be evaluated in such form. Therefore, we applied a DCT post-computing to obtain 39-dim. pseudo MFCC_D_A_Z (the 0-th coeff. could not be computed) features that are regularly comparable with the baseline reference. The algorithm itself operates with several tunable-free parameters: 1) the length of the stacking window (N), 2) the number of retained principal components (M) and 3) the type and amount of the clean speech prior data. Based on a deep analysis, we determined that $N = 20$ and $M = 40$ is the optimal configuration for our data. Note that, we chose $d = 26$ mel-filters according to the sampling frequency $f_s = 16$kHz. Initially, we randomly selected 500 recordings from the clean training set "Court Room".

We achieved WER equal to 75.98% with the mentioned setup. In order to find the best configuration of the algorithm for the following experiments, we explored different prior speech data with fixed parameters M and N. The obtained results are listed in Tab. 2 with other notable information. All prior data were acquired from our particular Slovak corpora from different domains (judicature, CISCO course, broadcast news (Viszlay et al., 2016)). From the table we can observe that the best result was achieved, when recordings of CISCO lectures were employed as the speech prior. These data were the cleanest, without any noises or music in the background. Moreover, we found that the quality of the prior data in sense of clearness (unlike the amount) is essential for the successful dereverberation.

7.3. Performance and influence of global PCA

In the next stage, we replaced the DCT by PCA in the post-computing procedure. The PCA matrix was trained on the clean training data, as is typical in ASR. The 26-dim. input features were reduced to 13 and expanded with Δ and Δ^2 coeffs. It is known that PCA has better decorrelation effect that is proven in Tab. 3 (part I.), where the dereverberation error rate was reduced to 62.98%. The global PCA was applied in context-free manner and is independent from that applied in the dereverberation process.

7.4. Multi-condition training (MCT)

The first phase of MCT is evaluated in Tab. 3 (part II.). According to the mentioned facts about the MCT data (see Section 5.), the REV dataset represents the artificially reverberated training set, the ENH dataset corresponds to the enhanced REV dataset and the notation CLEAN2ENH stands for the pseudo-clean dataset. From the results we can deduce that the highest multi-conditionality is ensured, when all three datasets are joined together ($3 \times 80 = 240$ h) thus the acoustic-condition mismatch is minimized as much as possible. The WERs are compared to the previous stage (72.50%), without PCA so far. Utilizing the MCT strategy, we were able to reduce the WER by -4.43% compared to feature enhancement.

7.5. System fusion and fine tuning

After proving that PCA brings a significant improvement and that MCT also yields an appreciable improvement, we decided to fuse both systems together. The success of the fusion can be seen in Tab. 3 (part III.), where WER=60.65% was achieved by applying PCA to the MCT datasets. After optimization, we obtained WER=58.35% at 51 dimensions ($17 + \Delta + \Delta^2$). To decrease the error even more, we carried out language model adaptation (LMA) to the specific topic that resulted in a modest error decrease.

We employed CD-2DLDA in the last evaluation stage in the hope that the PCA features can be improved by discriminative transformation. This hypothesis was proven to be true and its positive effect is demonstrated in line denoted as CD-2DLDA*, where '*' denotes the 'CLEAN+ENH+CLEAN2ENH (51)' dataset. The CD-2DLDA transform was set to $L = 6$, $R = 4$ thus the output features had 48-dims. $(24 + \Delta)$. At last, we generated an adapted LM for the CD-2DLDA system but it did not succeed at all (CD-2DLDA* (48)+LMA). Therefore, we concluded that the WER level was reduced for the presented system as much as possible.

I. Features (39-dim.)	WER (%)
ENH (FBANK+DCT)	72.50
ENH (FBANK+PCA)	**62.98**
II. MCT	**WER(%)**
CLEAN+REV	69.99
CLEAN+ENH	68.67
CLEAN+REV+ENH	69.33
CLEAN+ENH+CLEAN2ENH	**68.07**
III. MCT + PCA (+CD-2DLDA)	**WER(%)**
CLEAN+ENH+CLEAN2ENH (39)	60.65
CLEAN+ENH+CLEAN2ENH (51)	58.35
CLEAN+ENH+CLEAN2ENH (51)+LMA	**58.14**
CD-2DLDA* (48)	**57.79**
CD-2DLDA* (48)+LMA	**57.79**

Table 3: Influence of global PCA (I.), Comparison of multi-condition datasets (II.) with subsequent PCA and CD-2DLDA trained multi-conditionally (III.)

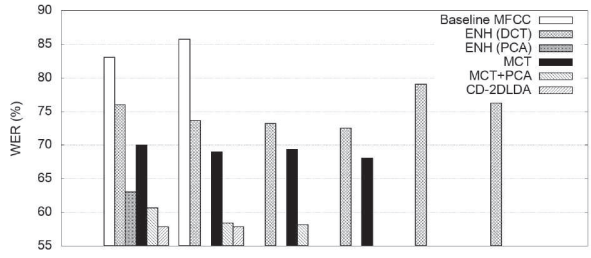

Figure 4: Overall comparison of the presented approaches

Finally, we give a summary in Fig. 4 through overall evaluation and comparison of the presented approaches. It can be concluded that the reverberation method itself reduces the reference level of WER by -10.53%. The global PCA is shown to be very effective (-9.52%), when it is applied instead of DCT as the post-computing step. We prove that the MCT framework fused together with linear transformations helped to reduce the condition mismatch considerably, exactly by -14.71%. Ultimately, the total absolute decrease achieved with respect to the baseline reference is -24.88%.

8. Conclusions and future intentions

In this paper, we reported an application result of building a reverberant-robust Slovak LVCSR system that uses a microphone array to capture speech in a meeting room. Several methods and techniques are employed to address the problem of the reverberant speech. We are aware that the presented system is not currently able to suppress the great impact of such high-level reverberation but the achieved absolute decrease of WER is quite interesting.

Our nearest future intentions are focused on replacing the conventional acoustic modeling and speech recognition approach by neural network based techniques using the Kaldi toolkit, definitely.

Acknowledgments

The research presented in this paper was supported by the Slovak Research and Development Agency under the project APVV-15-0517 and by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research project VEGA 1/0511/17.

9. References

Brandstein, M. S. and D. B. Ward, 2001. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag, New York.

Darjaa, S., M. Cerňak, M. Trnka, M. Rusko, and R. Sabo, 2011. Effective triphone mapping for acoustic modeling in speech recognition. In *Proc. of INTERSPEECH*. Florence, Italy.

Dennis, J. and T. H. Dat, 2015. Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: I2R'S system description for the ASpIRE challenge. In *Proc. of ASRU*. Scottsdale, AZ, USA.

Garofolo, J. S. et al., 1993. TIMIT Acoustic-phonetic continuous speech corpus. Linguistic Data Consortium.

Harper, M., 2015. The Automatic Speech recogition In Reverberant Environments (ASpIRE) challenge. In *Proc. of ASRU*. Scottsdale, AZ, USA.

Hiľovský, M., J. Greššák, M. Lojka, and J. Juhár, 2016. MAPL - Microphone Array Processing Library. In *Proc. of the International Symposium ELMAR*. Zadar, Croatia.

Jolliffe, I. T., 1986. *Principal Component Analysis*. New York, USA: Springer-Verlag.

Kalinli, O., M. L. Seltzer, J. Droppo, and A. Acero, 2010. Noise adaptive training for robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):1889–1901.

Keronen, S., H. Kallasjoki, K. J. Palomaki, G. J. Brown, and J. F. Gemmeke, 2015. Feature enhancement of reverberant speech by distribution matching and nonnegative matrix factorization. *EURASIP Journal on Advances in Signal Processing*, 2015(76):1–14.

Kinoshita, K., M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, 2013. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, USA.

Le, Q. and T. Mikolov, 2014. Distributed representations of sentences and documents. In *Proc. of the International Conference on Machine Learning (ICML)*. Beijing, China.

Lee, A., T. Kawahara, and K. Shikano, 2001. Julius - An open source real-time large vocabulary recognition engine. In *Proc. of EUROSPEECH*. Aalborg, Denmark.

Lojka, M., S. Ondáš, M. Pleva, and J. Juhár, 2014. Multithread parallel speech recognition for mobile applications. *Journal of Electrical and Electronics Engineering*, 7(1):81–86.

Mitra, V., J. Van Hout, W. Wang, M. Graciarena, M. McLaren, H. Franco, and D. Vergyri, 2015. Improving robustness against reverberation for automatic speech recognition. In *Proc. of ASRU*. Scottsdale, AZ, USA.

Mitra, V., W. Wang, Y. Lei, A. Kathol, G. Sivaraman, and C. Y. Espy-Wilson, 2014. Robust features and system fusion for reverberation-robust speech recognition. In *Proc. of REVERB Workshop*. Florence, Italy.

Nadiri, O. and B. Rafaely, 2014. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(10):1494–1505.

Palomäki, K. J. and H. Kallasjoki, 2014. Reverberation robust speech recognition by matching distributions of spectrally and temporally decorrelated features. In *Proc. of REVERB Workshop*. Florence, Italy.

Rafaely, B., 2015. *Fundamentals of Spherical Array Processing*. Springer-Verlag, Berlin.

Rajnoha, M., 2009. Multi-condition training for unknown environment adaptation in robust ASR under real conditions. *ACTA POLYTECHNICA, Journal of Advanced Engineering*, 49(2):3–7.

Ribas, D., E. Vincent, and J. R. Calvo, 2015. Full multicondition training for robust i-vector based speaker recognition. In *Proc. of INTERSPEECH*. Dresden, Germany.

Staš, J., D. Hládek, and J. Juhár, 2017. Semantic indexing and document retrieval for personalized language modeling. In *Proc. of the International Symposium ELMAR*. Zadar, Croatia.

Staš, J., P. Viszlay, M. Lojka, T. Koctúr, D. Hládek, E. Kiktová, M. Pleva, and J. Juhár, 2015. Automatic subtitling system for transcription, archiving and indexing of Slovak audiovisual recordings. In *Proc. of the Language & Technology Conference (LTC)*. Poznań, Poland.

Stolcke, A., 2002. SRILM - An extensible language modeling toolkit. In *Proc. of ICSLP*. Denver, Colorado.

Viszlay, P., M. Lojka, and J. Juhár, 2014. Class-dependent two-dimensional linear discriminant analysis using two-pass recognition strategy. In *Proc. of EUSIPCO*. Lisbon, Portugal.

Viszlay, P., J. Staš, T. Koctúr, M. Lojka, and J. Juhár, 2016. An extension of the Slovak broadcast news corpus based on semi-automatic annotation. In *Proc. of LREC*. Portorož, Slovenia.

Wölfel, M. and J. McDonough, 2009. *Distant speech recognition*. John Wiley & Sons Ltd.

Ye, J., R. Janardan, and Q. Li, 2005. Two-dimensional linear discriminant analysis. *Advances in Neural Information Processing Systems*, 17:1569–1576.

Yoshioka, T., A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, 2012. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. *IEEE Signal Processing Magazine*, 29(6):114–126.

Young, S. et al., 2006. *The HTK Book (for HTK Version 3.4)*. Cambridge University.