Investigation and Development of Methods for Improving Robustness of Automatic Speech Recognition Algorithms in Complex Acoustic Environments

M.L. Korenevsky, Yu. N. Matveev and A.V. Yakovlev

Abstract Aims and objectives of the study are described; state-of-the-art techniques in the study area are outlined. Several effective approaches proposed in the study and targeted at robustness improvement in complex acoustic environments are described. They are multichannel alignment algorithm, vector Taylor series-based features compensation with phase-term modeling, and environment adaptation method based on GMM-derived features. Experimental results analysis and comparison to state of the art are presented.

Keywords Speech recognition · Robustness · Distortion · Noise Compensation · Adaptation · Beamforming · VTS · GMM-derived features

Introduction

The past decade has seen a rapid development of speech recognition technology, which has led to significant improvements in the recognition accuracy for all scenarios of its usage and the large introduction of recognition technologies into many spheres of human activity. The reasons for this development are primarily related to the widespread adoption of multilayer (deep) neural networks for acoustic modeling. In several tasks, this made it possible to closely approach to (or even exceed) the human level of recognition accuracy. However, under strong noises and acoustic distortions, especially nonstationary, automatic speech recognition (ASR) algorithms are still noticeably inferior to human abilities. Accuracy of the acoustic models which are almost error-free in recognition of clean speech,

M.L. Korenevsky (✉) · Yu.N. Matveev · A.V. Yakovlev
ITMO University, Saint Petersburg, Russia
e-mail: korenevsky@speechpro.com

Yu.N. Matveev
e-mail: matveev@speechpro.com

A.V. Yakovlev
e-mail: yakovlev@speechpro.com

K.V. Anisimov et al. (eds.), *Proceedings of the Scientific-Practical Conference "Research and Development - 2016"*, https://doi.org/10.1007/978-3-319-62870-7_2

generally deteriorates when they are used in complex acoustic environments (strong noises, distant microphone, reverberation, etc.), i.e., even state-of-the-art ASR systems are mostly not sufficiently robust. This paper is devoted to research and development of new approaches to improve the robustness of ASR algorithms with the emphasis on using of neural network acoustic models.

Aims and Objectives

The aims of this study were to design new methods and software/engineering solutions for real-time automatic continuous speech recognition in complex acoustic environments. The designed methods should provide: noise suppression in the processed speech signal with a minimal distortion of its spectrum and as a consequence, its intelligibility improvement; voice activity detector (VAD) reliability improvement; accounting and compensation of the influence of noisy conditions on the recognition accuracy; speech recognition improvement in acoustic conditions different from those used for acoustic models training.

The relevance of research directions is confirmed by the fact that still there is no commercially successful speech recognition product, which would provide human-comparable recognition accuracy in the complex acoustic environment.

State of the Art in Study Area

Research in improving the robustness of ASR algorithms have a long history, first significant studies date back to 1970s, see for example [1]. At that time the main direction of increasing speech recognition accuracy was to preprocess speech signal itself to remove or at least suppress a noise component strongly in order to improve speech quality and intelligibility. This direction is still being actively developed, although conventional denoising techniques based on signal processing methods are gradually superseded with more sophisticated approaches which use non-negative matrix factorization (NMF) [2], filtering based on spectral masks generated by neural networks [3], missed data restoration techniques [4], and so on. The special place in this direction belongs to the processing of signals from several microphones (microphone array) [5]. Such approaches make it possible to take into account geometric features of relative positions of microphones and speaker and to "beamform" microphone array in such a way to amplify speech signal from a target direction and to suppress interference and noise from all other directions. Development of such processing methods is especially important for the applications like "smart home" when microphones are located in several parts of the room and both location and orientation of speaker's head are unknown in advance. In order to promote the development of multi-microphone approaches in robust speech recognition, several international competitions like CHiME (Computational

Hearing in Multisource Environments) Challenge [6–9] have been organized in recent years.

One more direction in improving ASR system robustness is using robust acoustic features, i.e., those whose distribution is distorted only slightly on changes of acoustic conditions (and which still keep good abilities to discriminate speech phones). In developing such features, researchers often refer to human auditory system which is able to recognize speech even in very adverse conditions. The examples of robust acoustic features based on auditory system processing are PNCC (Power-Normalized Cepstral Coefficients) [10] or gammatone filterbanks energies [11].

A similar problem of features variability reduction is also solved by various normalization methods like CMVN (Cepstral Mean and Variance Normalization) [12] or more general histogram equalization [13].

Besides, a large group of developed approaches are aimed at not in increasing features resistance to different distortions but instead try to explicitly remove the influence of these distortions—these are feature compensation methods. The most noticeable techniques among them are SPLICE (stereo-piecewise linear compensation for environment) [14], which uses statistics of joint speech and noisy features distribution to construct piecewise linear transform from noisy to clean speech features, and VTS (Vector Taylor Series) [15] which uses approximate linearization of nonlinear model of speech distortion by noise and channel to construct similar transform.

The idea of VTS is also applied in other groups of approaches for robustness improving, namely in adaptation of acoustic models to acoustic environment changes, i.e., adjusting models trained for clean speech recognition to the acoustic features distortions. The same linearized distortion model is used to modify features distribution parameterized as a GMMs (Gaussian Mixture Model). A number of other successful approaches to adapt GMM-based acoustic models were also developed such as MLLR, CMLLR (fMLLR) [16], MAP [17], PMC [18], and some their combinations.

However, in the past years, GMM-HMM acoustic models were almost everywhere superseded by acoustic models based on deep neural networks (DNNs), which provide much better accuracy in the vast majority of tasks. DNNs need completely different ways of adaptation, development of which was in a very initial stage when our study started. In order to keep the network architecture most of the methods developed till that moment modified weights of a trained neural net by fine-tuning them on adaptation data with a backpropagation algorithm. This is rather computationally demanding and needs to create a copy of initial network (which has millions of parameters) for each new acoustic conditions.

During this study, new approaches were developed within three of above-mentioned robustness improvement directions. The architecture and programming implementation of experimental software were also developed, where these approaches were integrated into single speech processing and recognition pipeline for complex acoustic environments. In the following sections, these approaches and results of their application are considered in more detail.

Multichannel Alignment (MCA)

This algorithm first described in [19] is an adaptive microphone array (MA) beamforming method using an output of well-known Delay-and-Sum beamforming method [5] as a "reference" signal. The algorithm computes adaptive transfer functions for each microphone channel signals by means of «aligning» their spectra relative to that of the reference signal. Signals passed through the transfer functions are then averaged to provide the resulting speech signal. This approach makes the width of the MA's directivity pattern main lobe narrower (i.e., improves spatial directivity) and significantly reduces the level of sidelobes (i.e., suppresses noises and interferences received on them). The scheme of MCA processing (for the case of 4 microphones) is depicted in Fig. 1. It is worth noting that the reference signal may be presumably obtained from any other beamforming algorithm as well, and better the reference more noticeable the effect of MCA should be.

STFT and IFT stand for Short-Time Fourier Transform and Inverse Fourier Transform respectively, $D_i(f)$ are channels' delays (steering) vectors.

MCA algorithm was successfully applied in our submission to the CHiME Challenge 2015 where it has demonstrated competitive results compared to several well-known beamforming algorithms [20]. The important characteristics of MCA include the implementation simplicity, low computational complexity and resistance to target direction errors.

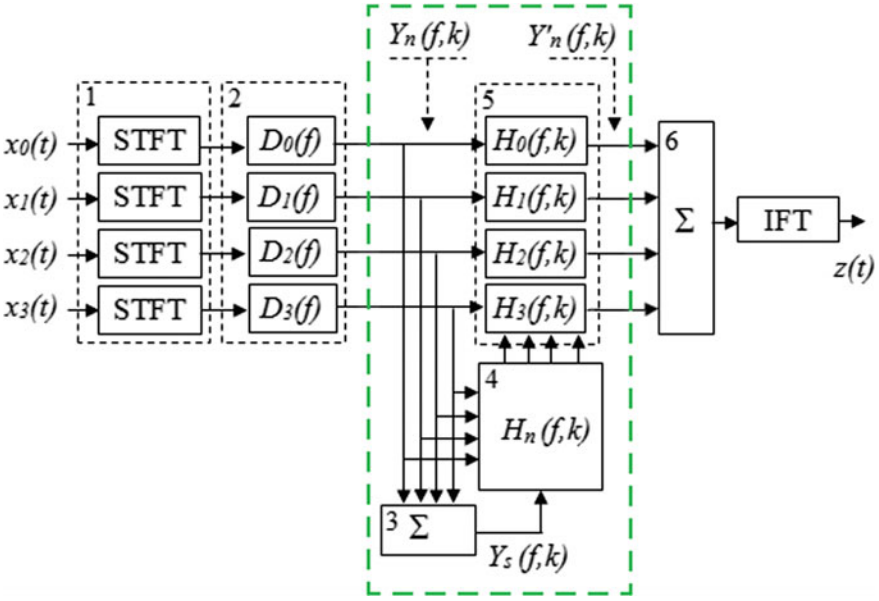

Fig. 1 Multichannel alignment algorithm

VTS Feature Compensation with a Phase-Term Modeling

VTS-based methods comprise an extremely important class of approaches to ASR robustness improvement: they are applied for features distortion compensation, adaptation and adaptive training acoustic models as well as for dealing with uncertainty remained after feature compensation during speech decoding. As it was already mentioned, VTS-based adaptation is applicable to only GMM-HMM acoustic models, therefore under widely used DNN-HMM framework its direct application is not possible. Thus, VTS-based acoustic features compensation (cleaning) becomes more important now.

The most widely used speech distortion model by noise and channel has the following form:

$y(t) = x(t) * h + n(t)$, where $x(t)$, $y(t)$, $n(t)$, and h denote clean speech, noisy speech, and noise signals as well as channel impulse response, respectively, and $*$ denotes convolution. When computing the most widespread MFCC (mel-frequency cepstral coefficients) features the signal is processed with several linear and non-linear transformations. This results in the following relation between the features of the above signals:

$$\mathbf{y} = \mathbf{x} + g(\mathbf{x}, \mathbf{h}, \mathbf{n}, \boldsymbol{\alpha}) = \mathbf{x} + \mathbf{h} + C\ \log\left(1 + e^{D(\mathbf{n}-\mathbf{x}-\mathbf{h})} + 2\boldsymbol{\alpha} \bullet e^{D(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}\right),$$

where $\mathbf{x}, \mathbf{y}, \mathbf{n}, \mathbf{h}$ denote vectors of MFCCs for clean speech, noisy speech, noise and channel response, $\boldsymbol{\alpha}$ is a "phase" vector, C and D—are the matrices of direct and inverse discrete cosine transform (DCT) and, finally, $\bullet$ is an elementwise (Hadamard) product of vectors. This model involves nonlinear vector-function g and its presence makes estimation of the clean speech features from noisy speech features extremely difficult. The essence of VTS method is a linearization of this nonlinearity by means of Vector Taylor expansion up to the first-order terms around some set of points:

$$\begin{aligned}\mathbf{y} \approx & \mathbf{x} + g(\mathbf{x}_0, \mathbf{h}_0, \mathbf{n}_0, \boldsymbol{\alpha}_0) + \nabla_{\mathbf{x}} g^T(\mathbf{x} - \mathbf{x}_0) + \nabla_{\mathbf{h}} g^T(\mathbf{h} - \mathbf{h}_0) \\ & + \nabla_{\mathbf{n}} g^T(\mathbf{n} - \mathbf{n}_0) + \nabla_{\boldsymbol{\alpha}} g^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0).\end{aligned}$$

This, of course, introduces some error into the model, but greatly facilitates the following inference.

The last nonlinearity term which contains phase vector $\boldsymbol{\alpha}$ was first taken into account in [21],[1] where it was demonstrated that this improves the model accuracy. However, in both just cited paper and subsequent ones, which use such distortion model phase vector was treated in some special ways: for example it was assumed to have equal components, which is not physically adequate, or its distribution

[1]This term was always discarded in previous papers as presumable being close to zero.

Table 1 Accuracy of the VTS in clean training scenario for different SNR values

SNR, dB	No VTS	VTS without phase	VTS with phase
Clean (>40)	99.08	99.01	99.04
20	94.32	98.22	98.47
15	84.65	96.67	97.49
10	64.07	92.57	94.52
5	36.46	82.47	86.79
0	16.18	56.71	64.58
−5	9.04	22.66	28.96

parameters were estimated from the trainset in advance and then considered as known. We proposed a new variant of VTS-based on the same model, where phase vector is treated as a multivariate Gaussian with the unknown parameters (as it is usually done for noise features vector **n**), and these parameters are inferred based on maximum likelihood principle and using EM-algorithm. This approach is not limited to noises available in the training set and does not put tight constraints on the phase vector structure. We derived EM expressions to update **n**,**α**, and **h** distributions parameters and formula for estimating clean speech features [22].

Experiments for assessing effectiveness of proposed VTS variant were performed, inter alia, on the Aurora2 database [23], which contains utterances of sequences of English digits distorted with different noises and channels. Obtained results, part of which is shown in Table 1, clearly demonstrate that the proposed method significantly improves the recognition accuracy compared to both VTS without phase-term modeling and especially to unprocessed noisy speech recognition.

Adaptation Based on GMM-Derived Features

It was already mentioned that well-designed adaptation methods for GMM-HMM acoustic models appeared to be not applicable after the migration to neural network acoustic models. Possibility of using GMM models for feature compensation gave rise to the idea of using GMM for adaptation as well if the features for DNN are not raw MFCCs but their GMM-based likelihoods of simple GMM-HMM acoustic model.

This idea led to the proposed method of adaptation based on GMM-derived features, designed in details in the papers [24–26]. The scheme of method application in its original variant (for speaker adaptation[2]) is depicted on Fig. 2.

[2]Method can be easily applied to environment but not speaker adaptation.

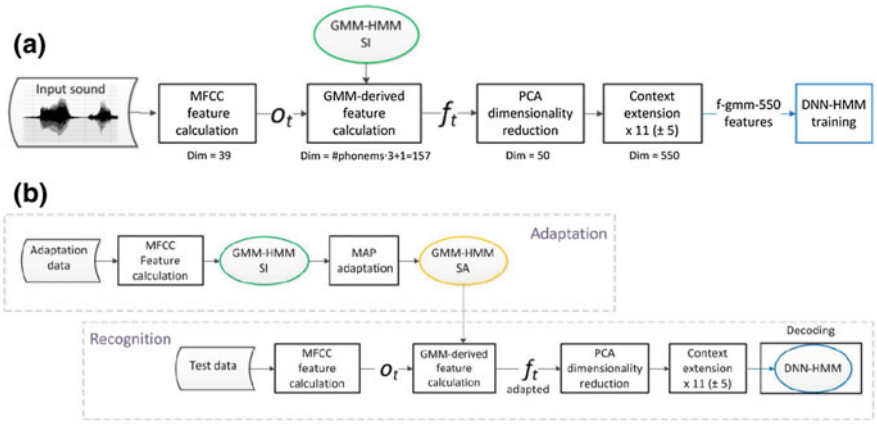

Fig. 2 Training and adaptation of DNN acoustic model based on GMM-derived features. **a** Training, **b** Adaptation and decoding

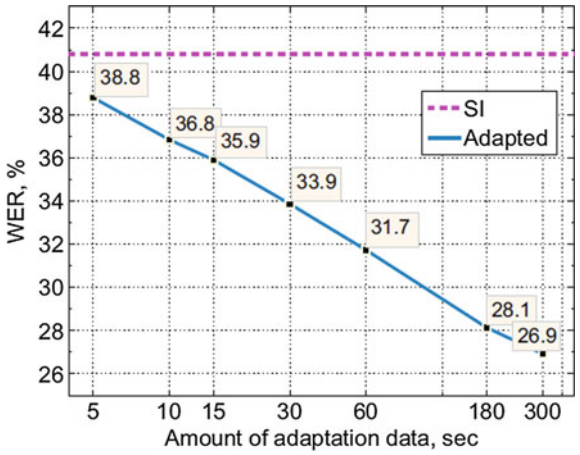

Fig. 3 Effect of adaptation depending on amount of adaptation data

As shown on the figure, speaker-independent (SI) GMM-HMM and DNN-HMM models are first trained. In this stage, the dimensionality of GMM likelihoods is reduced by PCA transform, and then features vector is extended by features from the neighboring frames. On the adaptation stage only GMM-HMM model is adapted (with MAP here) and then its (speaker adapted) outputs are fed into the same pipeline as for SI. As a result, the main DNN-HMM acoustic model remains unchanged. Number of experiments described in [24–26] show that although SI-DNN-HMM on GMM-derived features works worse than on MFCCs, its adaptation is extremely efficient. The illustration of the last statement (with application to speaker adaptation) is shown on Fig. 3.

Obviously, the described method may be also treated as a method of GMM-derived features compensation performed by means of GMM-HMM adaptation. Another variant of such compensation may be implemented based on the above-described VTS algorithm, where VTS-compensated MFCC features are fed into original GMM-HMM acoustic model to infer compensated GMM-derived features. We implemented the combination of both these approaches in the developed experimental software and found that they work well together.

Interestingly, the similar approach which combines VTS and GMM-derived features was recently considered in [27], however there VTS is used for direct GMM-HMM adaptation, therefore the direct comparison of these approaches is difficult.

Conclusions

Several different approaches which provide improvements of speech recognition accuracy in complex acoustic environments were developed in this study. They demonstrate competitive results on several well-known speech recognition benchmarks and have some advantages compared to many state-of-the-art analogues. The combination of the proposed methods is successfully implemented in the experimental software, which may be used as a basis for deployment of new ASD systems and devices, providing reliable speech recognition in adverse acoustic conditions.

Acknowledgments Research are carried out with the financial support of the state represented by the Ministry of Education and Science of the Russian Federation. Agreement no. 14.575.21.0033 27. June 2014. Unique project Identifier RFMEFI57514X0033.

References

1. Boll, S.F.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Tran. Acoust. Speech Sig. Process. **27**(2), 113–120 (1979)
2. Wilson, K.W, Raj, B., Smaragdis, P., Divakaran, A.: Speech denoising using nonnegative matrix factorization with priors. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, USA, 31 March–4 April 2008
3. Li, B., Sim, K.: An ideal hidden-activation mask for deep neural networks based noise-robust speech recognition. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 200–204, Florence, Italy, 4–9 May 2014
4. Raj, B., Seltzer, M., Stern, R.: Reconstruction of missing features for robust speech recognition. Speech Commun. **43**(4), 275–296 (2004)
5. Brandstein, M., Ward, D. (eds.): Microphone Arrays. Springer, Heidelberg (2001), 398 p
6. Barker, J., Vincent, E., Ma, N., Christensen, C., Green, P.: The PASCAL CHiME speech separation and recognition challenge. Comput. Speech Lang. **27**(3), 621–633 (2013)
7. Barker, J., Marxer, R., Vincent, E., Watanabe S.: The third 'chime' speech separation and recognition challenge: Dataset, task and baselines. Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2015)

8. Proceedings of the 4th International Workshop on Speech Processing in Everyday Environments (CHiME 2016), San Francisco, 13 Sept 2016, 90 p
9. Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M.: The second CHiME speech separation and recognition challenge: datasets, tasks and baselines. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, 26 May–31 May 2013 (2013)
10. Kim, C., Stern, R.M.: Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing (2012)
11. Shao, Y., Jin, Zh., Wang, D., Srinivasan, S.: An auditory-based feature for robust speech recognition. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, 19 Apr–24 Apr 2009
12. Viikki, O., Laurila, K.: Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Commun. **25**(1) (1998)
13. de la Torre, Á., Peinado, A.M., Segura, J.C., Pérez-Córdoba, J.L., Benítez, M.C., Rubio, A.J.: Histogram equalization of speech representation for robust speech recognition. IEEE Trans. Speech Audio Process. **13**(3) (2005)
14. Droppo, J., Deng L., Acero A.: Evaluation of SPLICE on the Aurora 2 and 3 tasks. Proceedings of the International Conference on Speech and Language Processing (ICSLP), pp. 29–32 (2002)
15. Moreno, P., Raj, B., Stern, R.: A vector taylor series approach for environment-independent speech recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),vol. 2, pp. 733–736, Atlanta, Georgia, USA, 7–10 May 1996
16. Gales, M.J.F.: Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**(2), 75–98 (1998)
17. Shinoda, K.: Speaker adaptation techniques for automatic speech recognition. Proceedings APSIPA ASC 2011 Xi'an (2011)
18. Gales, M.J.F., Young, S.J.: Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. **4**(5), 352–359 (1996)
19. Stolbov, M., Aleinik, S.: Speech enhancement with microphone array using frequency-domain alignment technique. Proceedings of 54-th International Conference on AES Audio Forensics, pp. 101–107, London (2014)
20. Prudnikov, A., Korenevsky, M., Aleinik, S.: Adaptive beamforming and adaptive training of dnn acoustic models for enhanced multichannel noisy speech recognition. Proceedings of 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015), pp. 401–408 (2015)
21. Deng, L., Droppo, J., Acero, A.: Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. IEEE Trans. Speech Audio Process. **12**(2), 133–143 (2004)
22. Korenevsky, M., Romanenko, A.: Feature space VTS with phase term modeling. Speech Comput. Lect. Notes Comput. Sci. **9811**, 312–320 (2016)
23. Hirsch, H., Pearce, D.: The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions. Proceedings of ISCA ITRWASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium (2000)
24. Tomashenko, N., Khokhlov, Y.: Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing. Proceedings of Interspeech, pp. 2997–3001 (2014)
25. Tomashenko, N., Khoklov, Yu.: GMM-derived features for effective unsuper-vised adaptation of deep neural network acoustic models. Proceedings of Interspeech (2015)
26. Tomashenko, N., Khokhlov, Yu., Estève, Y.: On the use of gaussian mixture model framework to improve speaker adaptation of deep neural network acoustic models. Proceedings of Interspeech (2016)
27. Kundu, S., Sim, K.C., Gales, M.: Incorporating a generative front-end layer to deep neural network for noise robust automatic speech recognition. Inter-speech (2016)