# Noise and Channel Normalized Cepstral Features for Far-speech Recognition

Michal Borsky, Petr Mizera, and Petr Pollak

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

**Abstract.** The paper analyses suitable features for distorted speech recognition. The aim is to explore the application of command ASR system when the speech is recorded with far-distance microphones with a possible strong additive and convolutory noise. The paper analyses feasible contribution of basic spectral subtraction coupled with cepstral mean normalization in minimizing of the influence of present distortion in such far-talk channel. The results are compared with reference close-talk speech recognition system. The results show the improvement in WER for channels with low or medium SNR. Using the combination of these basic techniques WERR of 55.6% was obtained for medium distance channel and WERR of 22.5% for far distance channel.

**Keywords:** distorted speech, far-speech recognition, cepstral features, spectral subtraction, cepstral mean normalization.

## 1 Introduction

The automatic speech recognition (ASR) systems have become a widely used assisting tools in the last decade [1]. The most frequent applications include online personal dictation systems [2], automatic broadcast transcription (subtitling) [3], [4], offline transcription of audio archive, key word spotting or finally systems for voice control of particular devices. The simplicity and convenience of voice interaction with the machines is a strong driving force for the research of deployment of its in office, household, car, or industry devices or machines. There are a lot of applications focused on replacing the current human-to-machine interfaces such as keyboard, mouse or touchpad.

Nowadays, very popular applications of voice driven interface is for a control of various devices or functionalities in so called smart-home [5]. The deployment of ASR system for voice command control in such applications requires a special tailoring at the levels of feature extraction and acoustic modelling as natural performance conditions of these systems are frequently rather adverse and they need to be to compensated because the requirement of speech input naturalness in smart-home environment leads to the usage of middle or far distance microphones, which are usually embedded in devices itself or in the walls or ceiling of the house and which disables the usage of directional microphones. When a microphone with omnidirectional characteristics is used, especially with far distance placement typically, it leads to the inevitable presence of various kinds of noise of rather high levels. Also an attenuation of speech collected

M. Železný et al. (Eds.): SPECOM 2013, LNAI 8113, pp. 241–248, 2013.

by a far microphone is rather high, so consequently, the resulting degradation of speech is really very high, and the accuracy of speech recognition falls down rapidly because standard features as MFCC or PLP are generally susceptible to strong noise [6] presence or to signal degradation.

Within this paper we would like to analyse a contribution and limitations of basic speech enhancement techniques and a possible impact on one-channel far-microphone speech recognition.

## 2 Far Channel Feature Extraction

As it is written above, middle- and far-speech recognition represent typically task with speech input extremely degraded by convolutory noise given mainly by reverberations, moreover, it may be strongly influenced by additive background noise. Consequently, some noise suppression techniques same as elimination of convolution distortion must be applied in such case.

There are various solutions of robust feature extraction working with signals of various level of degradation. Authors in [7] used the cepstral mean subtraction (CMS) to compensate for possible channel change in telephone band, when a sliding window of 400 frames was employed. It meant an averaging within approx. 4 seconds for their feature extraction setup of 25ms frame with 10ms shift. A similar approach to feature enhancement was implemented in [8] for three different kinds of noise (car, street noise, AWGN). A significant word error rate reduction (WERR) of 25.5% was reached for CAR noise.

Both the CMS and spectral subtraction (SS) were tested for robust speech recognition in [9]. The authors used the SS technique coupled with VAD to estimate noise power spectra obtained by averaging within speech pauses. The CMS algorithm was implemented by computing the long-time average of cepstral coefficient off-line and subtracted from all 13 coefficient but the zeroth. The pre-recorded noises were added to clean telephone signals at four various SNR levels. The results showed that SS could decrease the performance of the ASR system due to the introduction of non-linearities but testing database resulted in the WERR = 22.6 %. The mean word error rate reduction on the whole used database using only the CMS method was rather small, but the combination of SS and CMS yielded the WERR = 28.5 %.

*Noise cancellation in feature extraction*

There are various approaches for noise cancellation applied in ASR including sometimes quite sophisticated solutions using multichannel input [10]. But these techniques require much complex hardware same as higher computational cost. Due to these facts, they are usually not applied to increase the robustness of speech recognition but simple one-channel techniques are very popular and frequently used for these tasks.

Within our work we use SS technique described [6], [11] for the elimination of additive background noise. This technique was chosen because it works without need of voice activity detector. Within mentioned papers was also proved that it contributes reasonably to speech recognition in very noisy environments. It suppresses non-stationary noise when its spectral characteristics change slowly then speech ones. We can suppose near the same conditions in our approach in smart-home application.

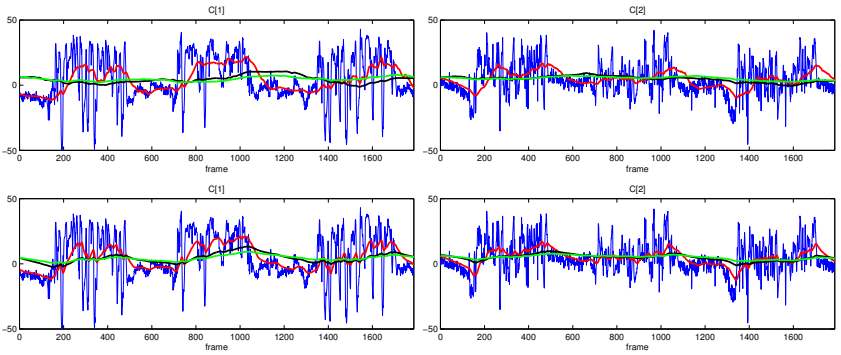


**Fig. 1.** CMS with cepstral mean estimation as EA/MA and smoothing time 1 s (red), 5 s (black), 10 s (green)

*Convolution distortion normalization*

Cepstral mean subtraction is the technique known already for several decades [12] or [13] and it is also used in world-wide spread tools for ASR such as HTK Toolkit [14]. General principle is clear and simple, however, the practical implementation differs, e.g. within HTK Toolkit the average cepstrum is computed only over whole utterance. It yields to various number of samples over which the average is estimated. Another drawback is the possibility to apply some approaches of CMS only in off-line mode.

We analyse two approaches of CMS, available now in [15], which can be easily implemented in on-line system. Firstly, it is standard computation of moving average (MA) over the long-time window of given length. The second approach is computation on the basis recursive exponential averaging (EA).

The key question is about a length of long-time window above which an average is compute. Particular authors work with various lengths of this window from 1 s up to values above 10 s. Fig. 1 illustrates averaging results for both solutions and it is clear that this window should be longer than 1 s, on the other hand from a value around 5 s the results starts being near the same.

## 3 Experiments

As mentioned above, the purpose of this study is to analyse the contribution and possible limitations of this basic techniques in the task of one-channel far-speech recognition. As a model of this situation data from Czech SPEECON database were used. Same utterance were here recorded simultaneously by several microphones located in different positions [16]. Recording conditions of these channels can be described by estimated values of SNR. This information for all channels CS0-CS3 is available in this database in particular annotation files. Statistics and distributions of speech SNR within this database are summarized illustratively in Fig. 2. More than 20 dB difference in SNRs between close and far distance speech proves that far channel data has significantly worse quality and that they well represent far distance input in smart-home application.

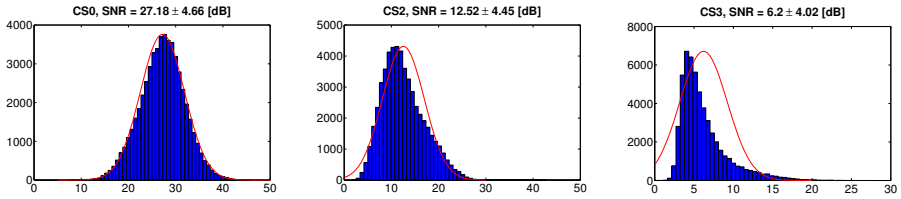


**Fig. 2.** SNR [dB] distribution in channels with estimated Gaussian fitting

*Robust feature implementation*

Above described robust features were computed using the tool CtuCopy [17] which offers many various strategies of parametrization in combination with additive noise suppression techniques, also SS technique chosen by us. Also both above described methods of convolution distortion normalization on the basis of CMS was additionally implemented into this tool. CtuCopy enables batch processing of more files (similarly as known HCopy tool [14]), however, the average was carried over signal boundaries. The last version of CtuCopy containing already described approaches of CMS computation is available for public usage.

Finally, we worked with MFCC were computed using the Ctucopy tool [17] with the following setup:

- 12 cepstral coefficients with c[0],
- 25 ms frame length with 15 ms step,
- 30 filters in full band $0 \div 8000$,
- static, dynamic, and acceleration features used.

Noise cancellation based on spectral subtraction was implemented with the following parameters:

- method extended spectral subtraction,
- spectra of the noise estimated in each frame with no crossover,
- integration constant p = 0.95,
- realized in magnitude domain,
- SS used before the application of the filter bank.

Cepstral mean subtraction was applied using both approaches, i.e. block and exponential averaging. Equivalent time constants for both methods were set to 1, 5, and 10 s. Commonly with SS, 14 different feature extraction setups which are summarized in the Tab. 1 were analysed.

**Table 1.** Parametrizations summary

| Param. | SS | T [s] | | |
|---|---|---|---|---|
| *mfcc* | no | - | | |
| *mfcc_ss* | yes | - | | |
| *mfcc_b/mfcc_exp* | no | 1 | 5 | 10 |
| *mfcc_ss_b/mfcc_ss_cms* | yes | 1 | 5 | 10 |

*Recognition task setup*

As the recognition task, small vocabulary recognition of 468 different commands with impossible repetition was chosen in our experiments. The utterances had a single word or multiple words structure and they also contained possibly used commands for household appliances. The testing part consists from 19 speakers with an overall length of about 15 minutes.

Speaker independent *acoustic models* for analysed channels were trained with the same amount of data which was about 51 hours of speech from 190 speakers. Final acoustic models had the following parameters: 43 different monophones including silence and short-pause expanded into tied-state cross-word triphones, 14 mixtures, static, dynamic, and acceleration features in 1 stream.

*Results*

Since all of the signals were recorded simultaneously using different microphones, channel distortion could be quantified basically by Euclidean cepstral distance computed between the reference CS0 signal and CS2/CS3 signal computed either from complete cepstral vector with coefficient $c_0$ (*CD0*) or just from the coefficients $c_1 \div c_L$ (*CD1*).

Tab. 2 shows results estimated from subset of approx. 2000 utterances from office part of SPEECON database. The trend observed for both CMS methods was the decrease in the (*CD0*) and (*CD1*) as the averaging time windowed increased in length. The (*CD1*) distance was consistently lower for independent CMS system than for the combined system, regardless of the channel. The differences were however very small.

The first experiments compares the results of a system without any noise suppression and a system with either SS or CMS for all channels. The application of standalone SS increased the robustness only in the case of CS3 channel. In both the CS0 and CS2 channels, the additive noise from the background in rather small, $SNR_{CS0}$ = 27.18 dB and $SNR_{CS2}$ = 12.52 dB. The induction of non-linearities and musical tones degraded the speech quality, which resulted in the increase of $WER$.

In the second experiment the accuracy was tested for a system with standalone CMS. In this case a clear improvement was reached for all setups on the CS2 channel, while

**Table 2.** Cepstral Euclidean distance for various parametrizations

| | CS2 | | CS3 | |
|---|---|---|---|---|
| | CD0 | CD1 | CD0 | CD1 |
| CS0 × CS_x | $\mu \pm \sigma$ | $\mu \pm \sigma$ | $\mu \pm \sigma$ | $\mu \pm \sigma$ |
| *mfcc* × *mfcc* | 42.57±*11.11* | 37.57±*11.30* | 54.91±*16.33* | 49.11±*18.87* |
| *mfcc* × *mfcc_SS* | 41.99±*11.32* | 38.79±*11.52* | 54.69±*17.04* | 50.36±*18.70* |
| *mfcc* × *mfcc_exp1* | 48.23±*14.02* | 43.44±*15.46* | 54.25±*16.80* | 48.26±*19.57* |
| *mfcc* × *mfcc_exp5* | 45.82±*12.45* | 41.06±*13.14* | 52.93±*15.74* | 46.93±*18.29* |
| *mfcc* × *mfcc_exp10* | 45.27±*12.14* | 40.48±*12.71* | 52.60±*15.55* | 46.57±*18.07* |
| *mfcc* × *mfcc_b1* | 49.25±*14.26* | 44.59±*15.62* | 55.02±*17.01* | 49.15±*19.71* |
| *mfcc* × *mfcc_b5* | 46.07±*12.39* | 41.41±*12.84* | 53.10±*15.52* | 47.18±*17.92* |
| *mfcc* × *mfcc_b10* | 45.51±*12.15* | 40.77±*12.67* | 52.77±*15.55* | 46.78±*18.03* |

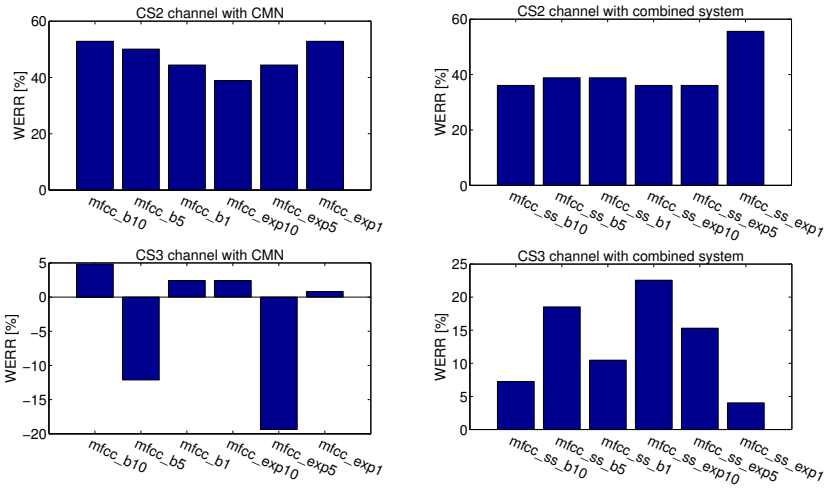


**Fig. 3.** WERR [%] for various parametrizations

**Table 3.** Reference and standalone CMS

| | CS0 | | CS2 | | CS3 | |
|---|---|---|---|---|---|---|
| Param | WER [%] | WERR [%] | WER [%] | WERR [%] | WER [%] | WERR [%] |
| *mfcc* | 1.89 | *0* | 9.73 | *0* | 33.51 | *0* |
| *mfcc_b10* | 2.43 | *-28.57* | 4.59 | *52.86* | 31.89 | *4.83* |
| *mfcc_b5* | 2.43 | *-28.57* | 4.86 | *50.05* | 37.57 | *-12.11* |
| *mfcc_b1* | 2.16 | *-14.28* | 5.41 | *44.39* | 32.7 | *2.41* |
| *mfcc_exp10* | 2.97 | *-57.14* | 5.95 | *38.84* | 32.7 | *2.41* |
| *mfcc_exp5* | 2.16 | *-14.28* | 5.41 | *44.39* | 40 | *-19.36* |
| *mfcc_exp1* | 2.16 | *-14.28* | 4.59 | *52.82* | 33.24 | *0.8* |

**Table 4.** SS and combined system

| | CS0 | | CS2 | | CS3 | |
|---|---|---|---|---|---|---|
| Param | WER [%] | WERR [%] | WER [%] | WERR [%] | WER [%] | WERR [%] |
| *mfcc_ss* | 2.43 | *-28.57* | 12.7 | *-30.52* | 29.46 | *12.08* |
| *mfcc_ss_b10* | 2.43 | *-28.57* | 6.22 | *36.07* | 31.08 | *7.25* |
| *mfcc_ss_b5* | 2.16 | *-14.28* | 5.95 | *38.84* | 27.3 | *18.53* |
| *mfcc_ss_b1* | 2.43 | *-28.57* | 5.95 | *38.84* | 30 | *10.47* |
| *mfcc_ss_exp10* | 3.24 | *-71.42* | 6.22 | *36.07* | 25.95 | *22.56* |
| *mfcc_ss_exp5* | 1.62 | *14.28* | 6.22 | *36.07* | 28.38 | *15.3* |
| *mfcc_ss_exp1* | 1.89 | *0* | 4.32 | *55.60* | 32.16 | *4.02* |

the CS0 showed the degradation in accuracy. The results for CS3 channel were mixed. The time constant of 5 seconds for both averaging methods proved to be unfit. The EA/MA methods with time constant 1/10 seconds performed better and increased the accuracy when compared to standard feature extraction.

In the last experiment the combination of both methods was tested. The combined system proved to be the most effective when for both noisy channel an improvement was reached for any setup and even a slight decrease of 0.27% in $WER$ for CS0 channel was observed.

## 4 Conclusions

The analysis of basic cepstral features applicable into far-talk speech recognition has been realized within this paper. Various setups of cepstral mean subtraction completed by extended spectral subtraction have been tested on the middle- to far-distance microphone recordings and compared to reference headset microphone recordings. The contribution of CMS was overall positive for CS2 channel, while for the the CS3 channel the time constant of 5 seconds proved to worse the accuracy for both block and exponential averaging. The WERR for *mfcc_exp1/10* and *mfcc_block1/10* was up to 5%. In each case the usage of CMS for far-talk channel is necessary. For more noisy far-talk channel (CS3), the contribution of SS was also evident due to significantly lower SNR in this channel. This combination of SS and CMS achieved the best results WER decreased for all CMS setups. Exactly, for CS2 channel the highest WERR = 55.6% was obtained for *mfcc_ss_exp1*, for CS3 channel the highest WERR = 22.5% was obtained for *mfcc_ss_exp10*.

**Acknowledgements.** Research in this paper was supported by grant SGS12/143/OHK3/2T/13 “Algorithms and Hardware Realizations of Digital Signal Processing”.

## References

1. Ircing, P., Krbec, P., Hajic, J., Psutka, J., Khudanpur, S., Jelinek, F., Byrne, W.: On large vocabulary continuous speech recognition of highly inflectional language - Czech. In: INTERSPEECH, pp. 487–490 (2001)
2. Newton Media: Newton Dictate Home page (2013), `http://www.diktovani.cz`
3. Nouza, J., Žďánský, J., David, P.: Fully Automated Approach to Broadcast News Transcription in Czech Language. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 401–408. Springer, Heidelberg (2004)
4. Vaněk, J., Psutka, J.V.: Gender-dependent acoustic models fusion developed for automatic subtitling of parliament meetings broadcasted by the czech TV. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 431–438. Springer, Heidelberg (2010)
5. Chaloupka, J., Nouza, J., Zdansky, J., Cerva, P., Silovsky, J., Kroul, M.: Voice Technology Applied for Building a Prototype Smart Room. In: Esposito, A., Hussain, A., Marinaro, M., Martone, R. (eds.) Multimodal Signals. LNCS (LNAI), vol. 5398, pp. 104–111. Springer, Heidelberg (2009)

6. Rajnoha, J., Pollák, P.: ASR systems in noisy environment: Analysis and solutions for increasing noise robustness. Radioengineering 20(1), 74–84 (2011)
7. Nouza, J., Silovsky, J.: Fast keyword spotting in telephone speech. Radioengineering 18(4), 665–670 (2009)
8. Schuller, B., Wöllmer, M., Moosmayr, T., Rigoll, G.: Speech recognition in noisy environments using a switching linear dynamic model for feature enhancement. In: INTERSPEECH 2008, pp. 1789–1792 (2008)
9. Kermorvant, C.: A comparison of noise reduction techniques for robust speech recognition. Idiap-RR Idiap-RR-10-1999, IDIAP, IDIAP-RR 99-10 (1999)
10. Wang, L., Odani, K., Kai, A.: Evaluation of hands-free large vocabulary continuous speech recognition by blind dereverberation based on spectral subtraction by multi-channel LMS algorithm. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS, vol. 6836, pp. 131–138. Springer, Heidelberg (2011)
11. Sovka, P., Pollak, P., Kybic, J.: Extended spectral subtraction. In: EUSIPCO 1996, Trieste (September 1996)
12. Junqua, J.C., Haton, J.P.: Asr of noisy, stressed, and channel distorted speech. In: Robustness in Automatic Speech Recognition. The Kluwer International Series in Engineering and Computer Science, vol. 341, pp. 273–323. Springer, US (1996)
13. Droppo, J., Acero, A.: Environmental robustness. In: Springer Handbook of Speech Processing, pp. 653–680. Springer (2008)
14. Young, S., et al.: The HTK Book, Version 3.4.1, Cambridge (2009)
15. Fousek, P., Mizera, P., Pollak, P.: CtuCopy feature extraction tool (2013), `http://noel.feld.cvut.cz/speechlab/start.php?page=download&lang=en`
16. Pollák, P., Černocký, J.: Czech SPEECON adult database. Technical report (November 2003), `http://www.speechdat.org/speecon`
17. Boril, H., Fousek, P., Pollak, P.: Data-driven design of front-end filter bank for Lombard speech recognition. In: Proc. of Interspeech 2006, Pitssburgh (September 2006)