Comparison of Cepstral Normalization Techniques in Whispered Speech Recognition

Đorđe GROZDIĆ[1], Slobodan JOVIČIĆ[1], Dragana ŠUMARAC PAVLOVIĆ[1], Jovan GALIĆ[2], Branko MARKOVIĆ[3]
[1]*School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia*
[2]*Faculty of Electrical Engineering, University of Banja Luka, Patre 5, 78000 Banja Luka, Bosnia and Herzegovina*
[3]*Čačak Technical College, Svetog Save 65, 32000 Čačak, Serbia*
djordje.grozdic@sbb.rs

***Abstract*—This article presents an analysis of different cepstral normalization techniques in automatic recognition of whispered and bimodal speech (speech+whisper). In these experiments, conventional GMM-HMM speech recognizer was used as speaker-dependant automatic speech recognition system with special Whi-Spe corpus containing utterance recordings in normally phonated speech and whisper. The following normalization techniques were tested and compared: CMN (Cepstral Mean Normalization), CVN (Cepstral Variance Normalization), MVN (Cepstral Mean and Variance Normalization), CGN (Cepstral Gain Normalization) and quantile-based dynamic normalization techniques such as QCN and QCN-RASTA. The experimental results show to what extent each of these cepstral normalization techniques can improve whisper recognition accuracy in mismatched train/test scenario. The best result is obtained using CMN in combination with inverse filtering which provides an average 39.9 percent improvement in whisper recognition accuracy for all tested speakers.**

***Index Terms*—automatic speech recognition, cepstral analysis, hidden Markov models, speech analysis, whisper.**

I. INTRODUCTION

Verbal communication can be carried out through different speech modalities that can be generally classified into seven categories based on the modes of speech production: whispered speech, quiet speech, normal or neutral speech, expressive speech (such as affective and emotional speech), singing, loud speech (e.g. speech with Lombard effect) and shouted speech [1]. Due to a relatively frequent utilization of whisper in everyday communication, automatic whisper recognition becomes an ongoing research topic. Current research studies in speech technologies are more than ever before focused on atypical speech modes and their improvement in automatic speech recognition (ASR) [2-5]. However, so far there are only a few studies regarding whisper recognition [2], [6-10] and this paper presents one of them.

In contrast to other speech modes, whisper production is characterized by a lack of regular vibration at the vocal folds [11]. This absence of glottal vibrations affects and greatly degrades the performance of traditional ASR systems that are primarily designed for normally phonated speech. The lack of voicing and specific way of articulation during whispering causes differences in energy and spectral characteristics, by which whisper in contrast to other speech modalities drastically differs from the normal speech [2], [3], [12], [13]. These differences can also be easily observed from the statistical properties of cepstral distributions, such as mean, variance, kurtosis and skewness. Therefore, an occurrence of whisper alone or in combination with normal speech (bimodal speech production) impacts the accuracy of conventional ASR systems, which are built on statistically-based acoustic models. This negative impact is particularly noticeable in mismatched train/test scenarios [2], [3], [13], where ASR systems are trained on normal speech and tested with some other speech mode — in this case with whispered speech. However, in theory, this adverse effect of talking style variability can be partially reduced with the application of cepstral normalization. This study introduces for the first time comparison of different cepstral normalization techniques in automatic recognition of whispered speech. The experimental results demonstrate to what extent certain cepstral normalization technique can improve automatic recognition of isolated words when conventional GMM-HMM recognizer is tested with whisper and bimodal speech (whisper + normal speech) from Whi-Spe corpus [14]. The following cepstral normalization techniques are considered in this article:

1) *Cepstral Mean Normalization (CMN),*
2) *Cepstral Variance Normalization (CVN),*
3) *Mean and Variance Normalization (MVN),*
4) *Cepstral Gain Normalization (CGN),*
5) *Quantile-based Cepstral dynamics Normalization (QCN), and*
6) *Quantile-based Cepstral dynamics Normalization with RASTALP filtering (QCN-RASTALP).*

The results of these experiments point out that certain cepstral normalization techniques are better than others in reducing and alleviating differences between normal speech and whisper in mismatched train/test scenarios. This study as well introduces a new method, known as inverse filtering [13], that can be applied in combination with cepstral normalization techniques, thus enhancing further their performance. Namely, inverse filtering reduces spectral differences between normal speech and whisper, normalizes the shape of their cepstral distributions, and thus improves the effect of cepstral normalization. The best result is

This research work was supported in part by the Ministry of Education, Science and Technological development of the Republic of Serbia under grant numbers TR-32032 and OI-178027.

Digital Object Identifier 10.4316/AECE.2017.01004

obtained with the combination of CMN and inverse filtering which provides an average 39.9% improvement in whisper recognition accuracy for all tested speakers.

The reminder of this paper is organized as follows: The next section presents Whi-Spe corpus containing audio recordings of isolated words in normal speech and whisper that was used in the experiments of this paper. With examples from Whi-Spe corpus, Section III introduces the most important acoustical characteristics of whispered speech and its differences compared to normal speech as well as statistical properties of cepstral distributions in whispered and normal speech. Section IV presents several common and state-of-the-art cepstral normalization techniques. The experimental setup is described in Section V, followed by the results of word recognition accuracy in different train/test experiments and comparison of different cepstral normalization techniques. Section VI discusses additional improvement of cepstral normalization by inverse filtering. The last section comprises the conclusions.

II. WHI-SPE CORPUS

For research purposes of automatic whisper recognition, the special speech corpus entitled "Whi-Spe" (acronym of Whisper-Speech) was developed, containing audio recordings of isolated words in two speech modes — in normal speech and whisper. The database consists of a total of 10,000 audio recordings that were collected from 10 native Serbian speakers (5 male and 5 female) with proper articulation and correct hearing. Each speaker had read set of 50 phonetically balanced words ten times in both speech modes. In order to obtain comparable sound pressure levels (SPL) in normal speech and whisper, during whispering the microphone position was set close to the speaker's mouth (at distance of about 5 cm) while all along the normal speech recordings the microphone was at approximately 25 cm from the speaker's mouth. The recording sessions were carried out under quiet laboratory conditions in a sound booth with professional recording equipment and strict quality control procedure. The Whi-Spe is one of the few existing, systematically collected databases of parallel neutral and whispered speech [2], [6], [15], [16]. Most of them have a small or medium-sized vocabulary, while only some of them are transcribed and phonetically balanced. Moreover, the Whi-Spe corpus is publicly available for research purposes and allows future database upgrades. More pieces of information about Whi-Spe corpus can be found in [14].

III. ACOUSTICAL CHARACTERISTICS OF WHI-SPE CORPUS

As a result of different vocal tract shape and articulation organ's specific behavior, whisper has acoustical characteristics that are quite different from those of normal speech. These differences are noticeable both in time and frequency domain and the most important ones are: lower energy, flatter spectral slope and the shift in formants' locations [9]. Taking in mind the fact that whisper does not have glottal vibrations and thereby voicing, amplitudes of voiced phonemes (vowels at the first place) are considerably lower in whisper, while the amplitudes of unvoiced phonemes have similar intensity as in normal speech [2]. Because of these characteristics, whisper has significantly lower energy and more sensitive signal-to-noise ratio (SNR). However, under clear conditions with good SNR, the low energy level in whisper recordings can be simply amplified by closer microphone position. Nevertheless, the spectral differences between speech and whisper still remain. Due to the lack of periodic excitation and harmonic structure, whisper does not have fundamental frequency thus either the most prosodic features. Also, in spectral domain whisper is characterized by upward shifts of the first three formants, noisy structure and flatter spectral slope [17]. As an illustration, Fig. 1 shows long-term average speech spectrums (LTASS) of normal speech and whisper recordings from Whi-Spe database.

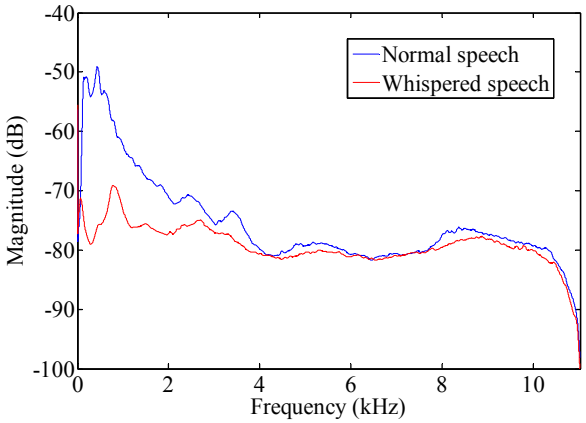

Figure 1. Long-term average speech spectrums (LTASS) of normal speech and whisper recordings from Whi-Spe database, showing the difference in spectral slope.

It is clear that normal speech has the much steeper spectral slope at frequencies up to 5 kHz, i.e. at the range where the first few formants and voicing are dominant. On the other side, whisper has almost flat spectrum. These spectral differences are the main cause of mismatch in train/test scenarios between normal speech and whisper, and they can be further analyzed in cepstral domain. Fig. 2 and Fig. 3 on the next page present cepstral distributions of the first two cepstral coefficients, c_0 and c_1, in normal speech and whisper.

When comparing distributions of c_0 in normal speech training data and whisper test data (Fig. 2), it can be seen that these distributions are centered around similar positions, but they do not ideally match. Since c_0 coefficients represent the energy of the speech signal, these distributions suggest that normal speech and whisper have similar energy level. Of course, this phenomenon is caused by a close microphone position during the whisper recording sessions (see reference [14]). Analysis of voiced and unvoiced speech segments from Whi-Spe database show that voiced sounds are dominant in normal speech (62.35%), and unvoiced sounds are dominant in whisper (99.2%). It is clear that in normal speech low energy components (lower c_0 values) correspond to unvoiced speech segments, and high energy components (higher c_0 values) represent voiced speech. In contrast to c_0 distributions, the difference between c_1 distributions in normal speech and whisper is much more noticeable (Fig. 3). In normal speech, distribution of c_1 coefficients is shifted to the right (higher

c_1 values) compared to c_1 distribution in whispered speech. In whisper, voiced sounds are not dominant, so distribution of c_1 coefficients has lower c_1 values. Based on the fact that c_1 is related to the spectral slope of the speech signal, it is clear that normal speech has steeper spectral slope than whisper.

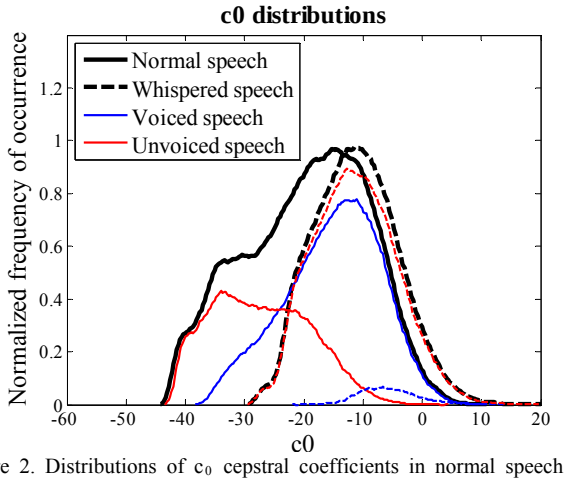

Figure 2. Distributions of c_0 cepstral coefficients in normal speech and whisper recordings from Whi-Spe database.

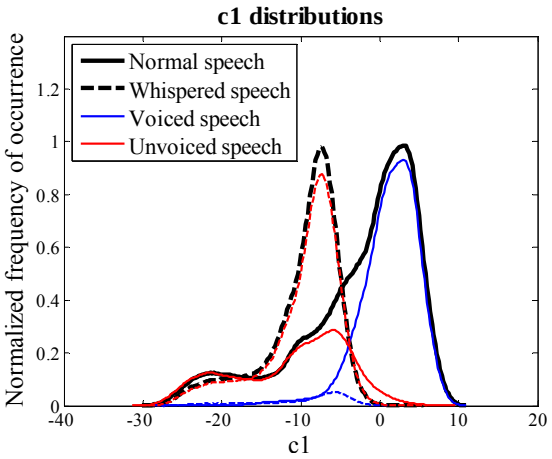

Figure 3. Distributions of c_1 cepstral coefficients in normal speech and whisper recordings from Whi-Spe database, showing a mismatch in their positions.

Since current ASR features are mostly in form of cepstral coefficients extracted from a short-time spectrum, especially spectral slope, formant and energy variations will directly impact ASR performance, introducing a degrading mismatch between input whisper data and neutral trained acoustic models [18]. However, this mismatch between cepstral distributions can be alleviated to a certain extent with the application of cepstral normalization. The following sections describe and compare several cepstral normalization techniques, and investigate how and to what extent these techniques can improve whisper recognition in mismatched train/test scenarios.

IV. Cepstral Normalization Techniques

The following cepstral normalization techniques are considered in this study:

A. CMN

Cepstral mean normalization (CMN) or cepstral mean subtraction (CMS) is widely used technique compensating for the speech signal variability in cepstral domain [19]. It is used as standard feature normalization technique for most large vocabulary ASR systems. The main focus of CMN is on convolutional distortions caused by characteristics of different communication channels or recording devices [20]. However, CMN is also partially effective in reducing the effects of additive environmental noise and talking style variability [20]. From this aspect, CMN could be useful in normalization of variability that appears in mismatched train/test scenarios with normal speech and whisper. In this study, CMN is applied in per-utterance fashion, meaning that cepstral mean, μ, is estimated from long time window (length of an utterance) and then subtracted from each cepstral sample:

$$c_{n,t}^{CMN} = c_{n,t} - \mu = c_{n,t} - \frac{1}{T}\sum_{t=1}^{T} c_{n,t}, \quad (1)$$

where n is the n-th cepstral dimension, and t is the index of cepstral sample in the window. All the following techniques in this paper are also applied in per-utterance fashion.

B. CVN

Cepstral variance normalization (CVN) [21] is popular supplement technique to CMN, that estimates variance, σ_n, of each cepstral dimension and normalizes it to unity:

$$c_{n,t}^{CVN} = \frac{c_{n,t}}{\sigma_n} = \frac{c_{n,t}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(c_{n,t}-\mu)^2}}, \quad (2)$$

where n is the n-th cepstral dimension, and t is the index of cepstral sample in the window.

C. MVN

CVN is often used in conjunction with CMN, where it contributes to robustness by scaling and limiting the range of deviation in cepstral features. This technique is known as mean and variance normalization MVN [21], and it is applied by the following equation:

$$c_{n,t}^{MVN} = \frac{c_{n,t}^{CMN}}{\sigma_n}, \quad (3)$$

where $c_{n,t}^{CMN}$ and σ_n are cepstral sample after CMN and standard deviation, respectively.

D. CGN

Recently proposed cepstral gain normalization (CGN) [21] showed better performances than CVN and MVN in noisy ASR tasks. CGN incorporates CMN and instead of variance estimates the sample dynamic range in each cepstral dimension directly from the maximum (c_{nmax}) and minimum (c_{nmin}) sample values and normalizes it to unity. It is calculated by the following equation:

$$c_{n,t}^{CGN} = \frac{c_{n,t}^{CMN}}{(c_{n\max} - c_{n\min})}. \quad (4)$$

E. QCN and QCN-RASTALP

Quantile-based cepstral dynamics normalization (QCN) is recently established normalization technique that reduces the impact of Lombard effect on ASR [5], [18]. Since some speech variations under Lombard effect are similar to whisper's variations, such as convolutional distortions

(spectral slope flattening, formant shifts and intensity changes) [18], it is expected that QCN will also have success in automatic whisper recognition. QCN method is inspired by older normalization techniques such as CMN, CVN, and CGN, whereby QCN estimates cepstral dynamic range using quantile intervals obtained from the sample histograms. In the following step, the histograms (cepstral distributions) are centered to the quantile mean and their variance is normalized to a unit inter-quantile interval:

$$c_{n,t}^{QCNj} = \frac{c_{n,t} - (q_j^{c_n} + q_{100-j}^{c_n})/2}{q_{100-j}^{c_n} - q_j^{c_n}}, \quad (5)$$

where q_j^{cn} and q_{100-j}^{cn} are *j*-th and (*100-j*)-th quantile estimates in the *n*-th cepstral dimension. In this way, QCN provides more accurate alignment of cepstral distributions in mismatched train/test scenarios in terms of their dynamic ranges. Another superiority to CMN, CVN, and CGN techniques is that QCN is more robust to different shapes of the sample distribution contours.

QCN can also be combined with a temporal filtering strategy RASTALP that is inspired by the popular RASTA filter. This combination is called QCN-RASTALP [5], [18] which additionally increases the robustness of cepstral coefficients to additive noise and reverberation.

V. Experimental Comparison of Cepstral Normalization Techniques in Whispered Speech Recognition

A. Experimental setup

Our experiments on isolated word recognition were performed using a conventional GMM-HMM speaker-dependent ASR system that was trained and tested with Whi-Spe database. As an input speech features, 13-dimensional Mel-frequency cepstral coefficients (MFCC) were used (including c_0), which were later normalized with different normalization techniques and applied in training and testing procedures. Feature extraction was performed using 24 ms window size with a frame shift of 8 ms, Hamming window and preemphasis coefficient of 0.97. Three different train/test scenarios were analyzed. In each scenario, GMM-HMM was trained with normal speech and then tested with: normal speech, whisper, and bimodal speech. Bimodal speech dataset was created by randomly taking 50% of normal speech recordings and 50% of whisper recordings from Whi-Spe corpus. The obtained results are presented in the following subsection.

B. Results

As expected, in matched train/test scenarios, i.e. when recognizer was trained and tested with normal speech, the performance of ASR was very high. The results of word recognition accuracy illustrated in Fig. 4 show that average word error rate (WER) ranges from 0.6% to 2.4% depending on a particular cepstral normalization technique that was applied. The smallest WERs are obtained with QCN and QCN-RASTA techniques, while the highest are noted in the case of CVN and MVN. It is also worth mentioning that depicted standard errors (SE) present standard deviation between results of different speaker-dependent systems divided by the square root of the sample size. All these values in matched train/test scenarios, including average WER and their SE, are small enough to be considered as commercially acceptable by terms of nowadays ASR standards.

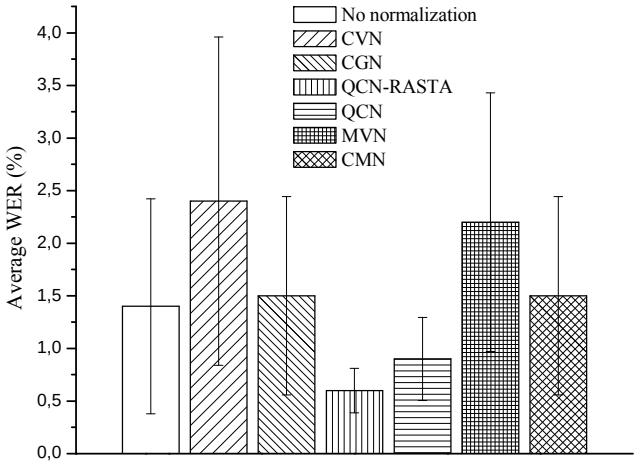

Figure 4. Average word error rate (WER) and standard error (SE) bars in matched train/test scenarios for all 10 test speakers using different cepstral normalization techniques.

However, in mismatched train/test scenarios, i.e. when recognizer was trained with normal speech and tested with a whisper, the performance of ASR drastically dropped down. The results of word recognition accuracy in this scenario are presented in Fig. 5, from which we can see the big increase in WER during whisper recognition. In the case when MFCC are applied without any normalization technique, the average WER is 87.3% for all tested speakers (this result is taken as a reference value for all the following comparisons of word recognition accuracy).

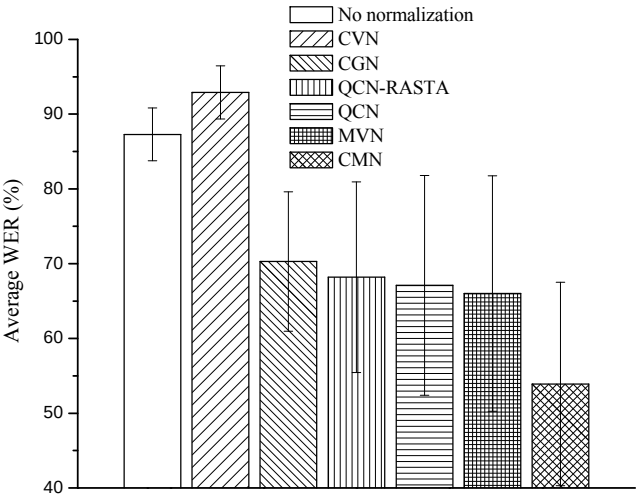

Figure 5. Average word error rate (WER) and standard error (SE) bars in mismatched train/test scenarios for all 10 test speakers using different cepstral normalization techniques.

Clearly, different bars in Fig. 5 illustrates that all normalization techniques improve word recognition rate in mismatched train/test scenarios except CVN. One more thing is important to be noticed. Beside WER, there is also obvious increase in standard deviation which suggests that there is some noticeable difference between the results of different speakers. This difference is caused by individual speaker's whispering style. Namely, there are two types of whisper based on vocal effort: hard whisper also known as forced whisper (high-energy whisper) and weak whisper

also called soft whisper (low-energy whisper) [22], [23]. Since high-energy whisper is by some characteristics more similar to normal speech than a low-energy whisper, it is obvious that speakers who tend to use forced whispering show better results in mismatched train/test scenarios. The average WER for all speakers in different test scenarios are presented in Table I.

TABLE I. PERFORMANCES OF DIFFERENT NORMALIZATION TECHNIQUES IN SPEECH, WHISPER AND BIMODAL SPEECH RECOGNITION

Normalization techniques	WER (%)		
	Speech	Whisper	(Speech +Whisper)
CMN	1.5	53.9***	27.7***
MVN	2.2	66.0**	34.1**
QCN	0.9*	67.1**	34.0**
QCN-RASTALP	0.6*	68.2**	34.4**
CGN	1.5	70.3**	35.9**
without normalization	1.4	87.3	44.4
CVN	2.4	92.9	47.7

*($p < 0.05$ *; $p < 0.01$ **; $p < 0.005$ ***; Confidence interval = 95%)*

As we can see, the best word recognition rate has CMN technique, which decreases WER from 87.3% to 53.9% and thus improves whisper recognition by 33.4%. Similarly, CMN proportionally improves recognition of bimodal speech and achieves 27.7% WER. On the other side, the only normalization technique that additionally degrades whisper recognition and thus bimodal speech is CVN. While variance represents well the sample dynamic range in the case of normal distribution, its accuracy reduces as the distribution skewness and shape deviate from normal [17], which happens in our case (see Fig. 3). MVN shows less recognition rate than CMN, because of applied variance normalization, and once again approves the statement that CVN has an adverse effect in whisper recognition. Although it is claimed that QCN and QCN-RASTALP are robust to different shapes of the sample distribution contours and variability of Lombard effect, these techniques show the lower performance in whisper recognition than CMN and MVN. QCN and QCN-RASTALP have 67.1% and 68.2% WER respectively, so they improve whisper recognition by 19% and bimodal speech by 10%. It is interesting that QCN-RASTALP technique additionally improves word recognition in matched train/test scenarios, and reduces WER to 0.6%. In the end, CGN technique has 70.3% WER and provides a modest improvement of 17% in whisper recognition and 8.5% in bimodal speech recognition.

The two-tailed Wilcoxon signed-rank test proves statistical significance of these improvements (see asterisks in Table I) and once again highlights that CMN technique is the one that best alleviates mismatch in speech/whisper scenario. The subtracting sample mean from the incoming test samples and thus removing slow varying cepstral component will assure that their dynamic range will match the one in the data used to train the ASR acoustic models. In this way, dynamic ranges of cepstral distributions in normal speech (training set) and whisper (test set) will be centered around the same mean value and the space between their positions depicted in Fig. 2 and Fig. 3 will be removed. However, the shape of cepstral distributions will still remain the same. The next section describes a new method based on inverse filtering which can reduce this difference in spectral shapes between whispered and normal speech and thus additionally improve cepstral normalization effect.

VI. FURTHER IMPROVEMENT OF CEPSTRAL NORMALIZATION BY INVERSE FILTERING

Having in mind the fact that cepstral normalization techniques are most effective when distributions' contours have symmetrical and Gaussian shape [18], this paper also analyses inverse filtering as a method that can additionally improve cepstral normalization techniques. Inverse filtering, also known as spectral whitening [13], flattens spectral slope, normalizes the shape of multimodal distributions in terms of their skewness and kurtosis and makes them more similar to a normal distribution. Fig. 6 illustrates LTASS of normal speech and whisper recordings from Whi-Spe database after inverse filtering.

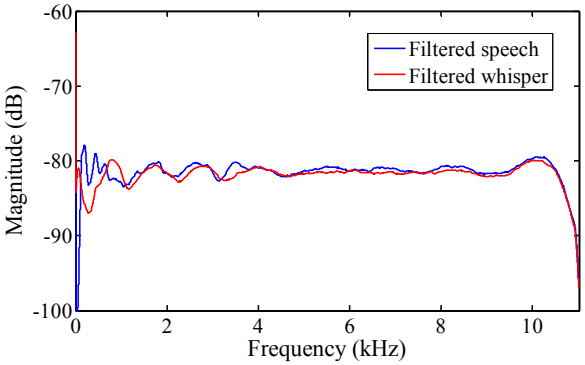

Figure 6. Long term average speech spectrums (LTASS) of normal speech and whisper recordings from Whi-Spe database after inverse filtering.

As it can be seen, inverse filtering suppressed voicing in speech as well as spectral slope, so the LTASSs of normal speech and whisper became more similar. However, these spectral changes didn't harm the formant structure, so there was no essential information loss. The impact of inverse filtering on cepstral distributions is presented in Fig. 7 and Fig. 8 on the next page. There are two important observations to be noticed. Firstly, due to voicing suppression, distribution of c_1 coefficients in normal speech is shifted to the left (lower c_1 values). The dynamic ranges and peaks of c_0 and c_1 distributions in both speech modes are now more precisely aligned. Secondly, multimodality of distributions is alleviated. The contours of distributions became more symmetrical, Gaussian shaped and similar to each other. After such modifications, MFCC features were again extracted, normalized and applied to GMM-HMM training. The new test results are presented in Table II.

TABLE II. PERFORMANCES OF DIFFERENT NORMALIZATION TECHNIQUES IN SPEECH, WHISPER AND BIMODAL SPEECH RECOGNITION AFTER INVERSE FILTERING

Normalization techniques	WER (%)		
	Speech	Whisper	(Speech +Whisper)
CMN	1.2	47.4	24.3
MVN	1.2	61.1	31.1
QCN	0.7	63.9	32.3
QCN-RASTALP	0.1	64.1	32.1
CGN	0.9	66.9	33.9
CVN	0.7	77.9	39.3

Inverse filtering enhanced all normalization techniques and improved accuracy in whisper and bimodal speech

recognition by average 5% and 2.5% respectively. Once again, CMN showed the best performance in whisper recognition (47.4% WER) and reduces WER by 39.9%.

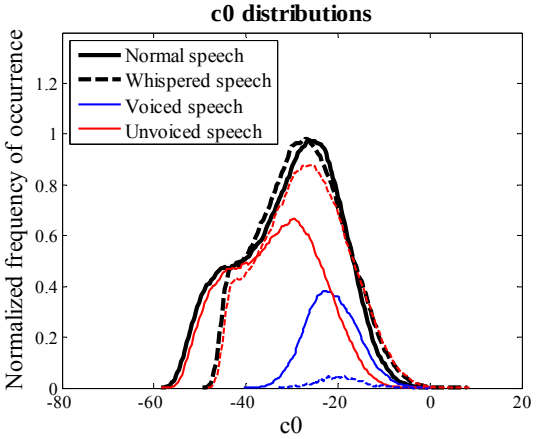

Figure 7. Distributions of c_0 cepstral coefficients in normal speech and whisper recordings from Whi-Spe database, demonstrating achieved alignment due to inverse filtering.

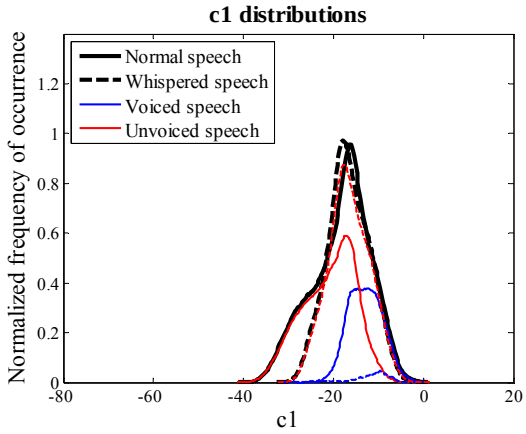

Figure 8. Distributions of c_1 cepstral coefficients in normal speech and whisper recordings from Whi-Spe database, demonstrating achieved alignment due to inverse filtering.

VII. Conclusion

Acoustical differences between normal speech and whisper are also reflected and noticeable in cepstral domain, primarily in terms of statistical properties of cepstral distributions, such as mean, variance, kurtosis, and skewness. This mismatch between cepstral distributions in normal speech and whisper directly impacts and degrades GMM-HMM performance in mismatched train/test scenarios. In order to examine whisper recognition in such situations, this paper compares different cepstral normalization techniques and suggests which one is the most suitable for whisper recognition. The results show that CMN is the best way to alleviate cepstral distributions between input whisper data and neutral trained acoustic models. Contrary, CVN degrades whisper recognition, and it is not recommended in mismatched train/test scenarios. Dynamic range normalization techniques and CGN also improve the accuracy of whisper recognition, but to a much lesser extent than CMN. Furthermore, additional improvement in performance is achieved by inverse filtering which reduces spectral differences between whisper and normal speech. Finally, the best whisper recognition accuracy is obtained with the combination of inverse filtering and CMN, resulting in 39.9% reduced WER.

References

[1] C. Zhang, J.H.L. Hansen, "Analysis and Classification of Speech Mode: Whisper through Shouted," in Proc. 8th Annu. Conf. Int. Speech Commun. Assoc. Interspeech 2007, Antwerp, 2007, pp. 2289-2292.

[2] T. Ito, K. Takeda, F. Itakura, "Analysis and recognition of whispered Speech," Speech Communication, vol. 45, pp. 139-152, Feb. 2005. doi:10.1016/j.specom.2003.10.005

[3] Đ.T. Grozdić, J. Galić, B. Marković, S.T. Jovičić, "Application of neural networks in whispered speech recognition," Telfor Journal, vol. 5, pp. 103-106, Nov. 2013.

[4] M.E. Ayadi, M.S. Kamel, F. Karray, "Survey on speech emotion recognition: Features, classification schemes and databases," Pattern Recognition, vol. 44, pp. 572-587, Mar. 2011.

[5] H. Boril, J.H.L. Hansen, "UT-Scope: Towards LVCSR under Lombard effect induced by varying types and levels of noisy background," in Proc. IEEE Int. Conf. Acoust. Speech Signal, ICASSP, Prague, 2011, pp. 4472-4475. doi:10.1109/ICASSP.2011.5947347

[6] S. Ghaffarzadegan, H. Boril, J.H.L. Hansen, "UT-Vocal Effort II: Analysis and constrained-lexicon recognition of whispered speech," in Proc. IEEE Int. Conf. Acoust. Speech Signal, ICASSP, Florence, Italy, 2014, pp. 2544–2548. doi:10.1109/ICASSP.2014.6854059

[7] A. Mathur, S.M. Reddy, R.M. Hegde, "Significance of parametric spectral ratio methods in detection and recognition of whispered speech." EURASIP J. Adv. Signal Process., pp. 157-177, Dec. 2012.

[8] C.Y. Yang, G. Brown, L. Lu, J. Yamagishi, S. King, "Noise-robust whispered speech recognition using a non-audible-murmur microphone with VTS compensation," in Proc. 8th International Symposium on Chinese Spoken Language Processing, ISCSLP, Hong Kong, China, 2012, pp. 220–223. doi:10.1109/ISCSLP.2012.6423522

[9] R.W. Morris, "Enhancement and recognition of whispered speech," Ph.D. dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, August 2003.

[10] Đ.T. Grozdić, S.T. Jovičić, M. Subotić, "Whispered speech recognition using deep denoising autoencoder," Engineering Applications of Artificial Intelligence, vol. 59, pp. 15-22, Mar. 2017. doi:10.1016/j.engappai.2016.12.012

[11] V.C. Tartter, "What's in a whisper?," Journal of the Acoustical Society of America, vol. 86, 1678-1683, 1989.

[12] B.P. Lim, "Computational differences between whispered and non-whispered speech." Ph.D. thesis, University of Illinois at Urbana-Champaign, 2011.

[13] Đ.T. Grozdić, S.T. Jovičić, J. Galić, B. Marković, "Application of inverse filtering in enhancement of whisper recognition," in Proc. 12th Symp. Neural Netw. Appl. Electr. Eng., NEUREL 2014, Belgrade, 2014, pp. 157-161. doi:10.1109/NEUREL.2014.7011492

[14] B. Marković, S.T. Jovičić, J. Galić, Đ.T. Grozdić, "Whispered Speech Database: Design, Processing and Application," in Proc. 16th International Conference, TSD 2013, Pilsen, 2013, pp. 591-598. doi:10.1007/978-3-642-40585-3_74

[15] P.X. Lee, D. Wee, H. Si, Y. Toh, B.P. Lim, N. Chen, B. Ma, V.J. College, "A whispered Mandarin corpus for speech technology applications," in Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH, Singapore, 2014, pp. 1598–1602.

[16] C. Zhang, J.H.L. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing." IEEE Trans. Audio, Speech Lang. Process. 19, 883–894, Aug. 2010. doi:10.1109/TASL.2010.2066967

[17] S.T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels," Acta Acust., vol. 84 (4), pp. 739-743, Jul. 1998.

[18] H. Boril, J.H.L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18 (6), pp. 1379–1393, Aug. 2010. doi:10.1109/TASL.2009.2034770

[19] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., vol. 55, pp. 1304–1312, 1974. doi:10.1121/1.1914702

[20] S.J. Hahm, H. Boril, A. Pongtep, J.H.L. Hansen, "Advanced Feature Normalization and Rapid Model Adaptation for Robust In-Vehicle Speech Recognition," in Proc. 6th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems, Seoul, 2013, pp. 14-17.

[21] S. Yoshizawa, N. Hayasaka, N. Wada, Y. Miyanaga, "Cepstral gain normalization for noise robust speech recognition," in Proc. IEEE Int. Conf. Acoust. Speech, Signal Process., Montreal, 2004, pp. 209-212.

[22] N.P. Solomon, G.N. McCall, M.W. Trosset, W.C. Gray, "Laryngeal configuration and constriction during two types of whispering," Journal of Speech and Hearing Research, vol. 32, pp. 161-174, Mar. 1989.

[23] P. Monoson, W.R. Zemlin, "Quantitative study of whisper," Folia Phoniatrica, vol. 36, pp. 53-65, 1984.