Effectiveness of Histogram Equalization and SyDOCC Features on Speech Recognition Performance on a Real-World Noisy Speech Task

Markus Müller, Martin Wagner, Juan Hussain, Sebastian Stüker, Alex Waibel
Karlsruhe Institute of Technology, Institute for Anthropomatics and Robotics, Interactive Systems Lab
Karlsruhe, Germany, Email: m.mueller@kit.edu

Abstract

When building systems for automatic speech recognition, one often faces the challenge of dealing with speech signals containing noise. This additional noise leads to a drop in recognition performance, especially, when the acoustic environment varies during training and testing of a system. There exist several approaches to deal with noisy data or mismatched conditions. We evaluate two different approaches: Histogram Equalization (HEQ) and Synchronized Damped Oscillator Cepstral Coefficients (SyDOCC). While HEQ tries to normalize the statistical properties of the input features in an unsupervised manner without requiring a noise estimate, SyDOCCs model the acoustic properties of the human ear more accurately than Mel-Frequency Cepstral Coefficients (MFCCs). We evaluate both approaches using data with artificially added noise as well as data that contains genuine noise due to the recording conditions.

Index Terms— *Noise-Robust Speech Recognition, Histogram Equalization, Synchronized Damped Oscillator Cepstral Coefficients*

Introduction

Large vocabulary continuous speech recognition (LVCSR) performs well under clean conditions. Those include a high signal-to-noise ratio (SNR) as well as matching channel characteristics. But when using data with a low SNR or mismatched conditions performance decreases. In recent years, a wide range of applications using LVCSR have arisen. In some of these applications the recording conditions cannot be controlled. While humans are able to understand speech even in such cases where the SNR is low or the channel characteristics are varying, automatic speech recognition does not perform well under such circumstances. Hence the need for robust speech recognition to deal with such issues.

There exist different approaches to deal with mismatched conditions. One possibility is to extract features that are inherently more robust or to post-process features to minimize the effect of different conditions.

Related Work

HEQ is an instance of a non-linear statistical matching algorithm which was derived from image processing. It was adopted for robust speaker recognition [1] and automatic speech recognition, e. g., [2, 3, 4]. Main implementation differences are the approximation of cumulative distribution functions (CDF) by order statistics [1], quantiles [5] or histograms [3, 4]. The reference distribution can be estimated from training data [3] or artificially chosen, e. g., as standard normal distribution [4]. HEQ can be regarded as non-linear improvement to the linear normalization methods Cepstral Mean Normalization (CMN) [6] and Cepstral Mean and Variance Normalization (CMVN) [7]. In addition to the first two statistic moments addressed by these methods, HEQ also normalizes all higher order statistic moments [4].

A traditional pre-processing pipeline features MFCCs. While they perform well under clean conditions, performance degrades significantly when dealing with mismatched conditions or noisy data. There exist several alternative input features that try to overcome the shortcomings of MFCCs. Those include RASTA-PLP [8] or PNCCs [9]. SyDOCCs [10] are a novel technique that try to mimic the properties of the human auditory system.

Methods for Noise-Robust Speech Recognition

In this paper, we evaluate two techniques to deal with noisy and/or mismatched acoustic conditions. They approach the problem at different stages of the preprocessing. While SyDOCCs are a new kind of input features, HEQ relies on post-processing features to level out mismatches.

Histogram Equalization

HEQ is an unsupervised, non-parametric normalization approach applied to an existing feature extraction pipeline. The goal is to reduce mismatch between feature spaces due to different acoustic conditions. Especially the cepstrum is particularly sensitve to additive noise, as it introduces non-linear distortions [11].

The HEQ transformation is applied independently for each feature vector component. Let $(x_1, \ldots, x_N)$ be the values of a feature vector component of a sequence of N feature vectors (e. g., of an utterance). A transformation $\mathrm{T}(x_i) = y_i$ is chosen to the effect that the distribution of the transformed sequence $(y_1, \ldots, y_N)$ matches a reference distribution, either gathered from training data, or simply a standard normal distribution [4]. This results in a monotone transformation which is non-linear in general. It can be found by utilizing the cumulative

distribution functions (CDF) of the input sequence, denoted as $\mathrm{CDF}_{\mathrm{data}}(x)$ and the targeted reference distribution, denoted as $\mathrm{CDF}_{\mathrm{ref}}(y)$. The transformed value y_i is found where the CDFs match.

$$\mathrm{CDF}_{\mathrm{data}}(x_i) \stackrel{!}{=} \mathrm{CDF}_{\mathrm{ref}}(y_i) \quad \Rightarrow \quad y_i = \mathrm{T}(x_i), \qquad (1)$$

and therefore

$$y_i = \mathrm{T}(x_i) = \mathrm{CDF}_{\mathrm{ref}}^{-1}(\mathrm{CDF}_{\mathrm{data}}(x_i)) \qquad (2)$$

as shown in [4].

When applying the transformation consistently for all sequences in training and testing, invariance to arbitrary non-linear distortions can be established. The main assumptions include: 1) For each feature vector component the distribution over a clean sequence is assumed to be the same for all clean sequences. 2) Each component can be normalized independent of each other. 3) The distortions caused by noise are monotone transformations and therefore fully invertible. As a main advantage of this approach no noise estimate or model is necessary.

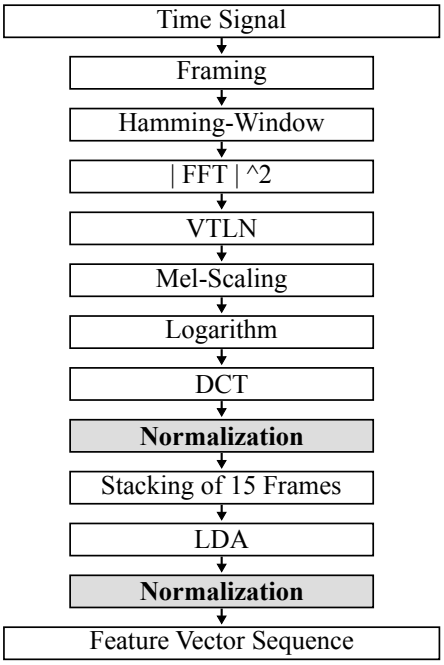

Figure 1: Utilized pre-processing pipeline for MFCCs: normalization is performed before and after LDA and may be either CMN, CMVN or HEQ. Normalization is omitted in the baseline system.

Synchronized Damped Oscillator Cepstral Coefficients

Speech recognition in noisy environments is difficult, as ASR systems are sensitive to changes in environmental conditions. The use of SyDOCCs [10] is a novel approach towards the pre-processing of an audio signal, motivated by the human auditory system. In contrast to MFCCs, SyDOCCs use a more sophisticated approach by mimicking the features of the inner ear. The auditory system acts as a frequency analyzer as well as a non-linear amplifier. In human sound perception the cochlea processes the sound waves with the help of hair cells [12]. Each hair cell reacts to a different resonance frequency. By modelling these characteristics, SyDOCCs aim at extracting more robust features that are insensitive to noise or mismatched conditions. This behaviour is modeled by using a Gammatone filter-bank in combination with damped oscillators. The filter-bank performs a frequency analysis and splits the signal into different frequency bins. The signal energy in each of these bins is used as a forcing function to excite oscillators. In contrast to MFCCs, SyDOCCs make use of the phase information. Each oscillator is synchronized to its two neighboring oscillators to oscillate in-phase. This raises the amplitude of the signal through the correlation of adjacent frequency bins. In case of uncorrelated noise, there is no phase alignment possible leading to no amplification. After the synchronization step, the envelope of the output of the oscillators is computed using modulation filtering which is followed by a power computation. As last step, a discrete cosine transform is applied. Out of this transform, only the first 13 coefficients were retained and based on them a joint feature vector using the 13 coefficients itself as well as Δs and $\Delta\Delta$s is constructed.

There are two degrees of freedom implementing SyDOCCs: The center frequency of the Gammatone filter-bank and the dampening factor of the oscillators. In our implementation, we used default parameters and did not vary them in the experiments.

Experiments and Results

We conducted our analysis using the Janus Recognition Toolkit (JRTk) [13] which features the IBIS decoder [14]. For each method we performed multiple sets of experiments. In one set of experiments we used clean speech recordings and artificially added noise. By doing so, we could assess the effectiveness of the methods for different SNR ratios. In a second set of experiments we used data featuring genuine noise to analyze the performance of the methods on a real-world task. We evaluated the systems by computing the word error rate (WER) on the output of the recognizer on a test set.

HEQ

We evaluated HEQ in a set of preliminary experiments using a corpus of English broadcast news. We artificially added recorded street noise. The noise was scaled to match different target SNRs in order to compare HEQ's performance to CMN, CMVN and a baseline system without normalization. Normalization methods are applied on an utterance basis both in training and testing. Their integration into the utilized pre-processing pipeline is shown in Figure 1. We use the standard normal distribution as reference distribution. The training data material has a total length of 187 hours, constituting of 7,336 speakers. In testing 3.66 hours were used, contain-

ing utterances by 79 speakers. The context dependent GMM-HMM recognizer utilizes 6,000 generalized quintphones, left-to-right tristate topology, 32 Gaussians per state, diagonal covariance matrices, trained by incremental splitting of Gaussians. In order to extract features from the input signal, we windowed the audio using a window length of 16ms and shifted that window over the data using a shift of 10ms.

Results are shown in Table 1 and Figure 2. HEQ outperforms other systems in the range of -5 and 20 dB but not for matched conditions. Experiments with different numbers of histogram bins did not show significant differences.

Table 1: Comparison of WERs from HEQ, CMN, CMVN and baseline for different SNRs.

SNR [dB]	WER			
	baseline	CMN	CMVN	HEQ
−20.0	96.7	96.0	**93.5**	94.2
−10.0	93.2	92.7	**88.0**	**88.0**
−5.0	87.3	86.2	77.1	**75.0**
0.0	73.5	70.8	61.0	**57.8**
5.0	55.6	51.5	46.6	**44.3**
10.0	41.5	39.0	37.3	**36.1**
20.0	31.4	30.0	30.1	**29.9**
30.0	29.6	**28.4**	28.8	28.9
clean	28.9	**27.9**	28.4	28.2

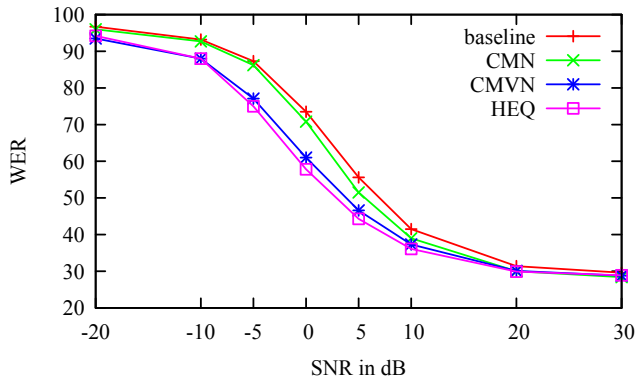

Figure 2: WER of HEQ compared to CMN, CMVN and baseline

In addition to experiments with artificially noised data, we also conducted a series of experiments using 8kHz narrow band telephone conversations containing genuine noise. We built systems for both Pashto (PUS) and Tagalog (TGL). Both systems are based on graphemes as acoustic units. In each of these experiments, we trained the systems using 100 hours of transcribed speech data. The test set consists of 10 hours. The pre-processing pipeline is identical to those of the previous experiment. The systems feature context-dependent models with 8,000 (PUS) or 10,000 (TGL) generalized quintphone states. Table 2 shows that we could not achieve a lower WER using HEQ as part of the pre-processing pipeline.

Table 2: Comparison of HEQ to baseline systems for data containing genuine noise.

System	Baseline	HEQ
PUS Grapheme	**69.1**	71.4
TGL Grapheme	**69.8**	72.8

SyDOCCs

Three sets of experiments were conducted to assess the performance of SyDOCCs. In the first experiment, the performance of an ASR system using SyDOCCs versus MFCCs was compared. We built a system for Italian broadcast news. 70 hours of audio were used for training. The test set consisted of 24 minutes of audio. Artificial white noise was added to the audio of the test set in different intensities. Our context dependent GMM-HMM recognizer uses 8,000 generalized quintphones. In contrast to our HEQ experiments, we use a window size of 25ms for the experiments with SyDOCCs as the original implementation uses that size. It was observed that the use of SyDOCCs on clean data leads to worse results compared to MFCCs, see Table 3. As the SNR decreases, using MFCCs results in greater losses in performance compared to SyDOCCs.

Table 3: Comparison of MFCCs and SyDOCCs using data with different SNRs of test data.

SNR [dB]	MFCC	SyDOCC	rel. gain
25.0	**27.0**	30.1	-10.3%
19.0	**29.0**	30.5	-4.9%
15.4	**30.9**	31.8	-2.8%
12.9	35.1	**32.6**	7.7%
11.0	41.4	**34.6**	19.7%
9.4	48.0	**36.7**	30.8%
8.1	55.0	**39.6**	38.9%
6.9	62.1	**43.8**	41.8%
5.9	70.6	**48.6**	45.3%
5.0	77.1	**52.9**	45.7%

In the second set of experiments we used the same recognizer as in the previous experiment, but instead of white noise we used street noise. We conducted 4 different experiments by selectively adding street noise to training and test set. We compared MFCCs with MFCCs and SyDOCCs combined. The results can be seen in Table 4. MFCCs show better results if the training data set is clean. Using noised data to train on, the combination of MFCCs and SyDOCCs outperform using MFCCs alone.

As final experiment, we assessed the performance on 8kHz telephone speech data containing genuine noise like we did for HEQ. The results are shown in Table 5. The baseline for a window size of 25ms is slightly worse compared to 16ms that we used for our HEQ experiments (Table 2). But like for HEQ, we did not see improvments from using SyDOCCs compared to MFCCs.

Table 4: Comparison of MFCC and **M**FCC + **S**yDOCC stacked with different combinations of clean and noised data sets

training	test	MFCC	M + S
clean	clean	**26.2**	28.3
clean	street noise	**36.3**	37.5
street noise	clean	27.7	**27.5**
street noise	street noise	36.4	**35.1**

Table 5: Comparison of **M**FCC + **S**yDOCC to baseline MFCC systems for data containing genuine noise.

System	MFCCs	M + S
PUS Grapheme	**71.5**	75.0
TGL Grapheme	**75.5**	77.9

Conclusion

We have implemented and evaluated two methods for dealing with speech degradations. Using data with artificial and genuine noise we have shown that both methods are capable of dealing with acoustic mismatches under certain conditions. As for the SyDOCCs, additional parameter tuning is required to achieve optimal performance: Varying the center frequency of the Gammatone filterbank or the dampening factor of the oscillators might have produced better results. In addition to that, both methods should be evaluated as an additional source of information with a LVCSR system featuring a DNN as part of the pre-processing pipeline.

Acknowledgement

The authors wish to thank Richard Stern for helpful discussions and the provided insight. This effort uses the IARPA Babel Program language collection releases IARPA-babel{104b-v0.4bY,106-v0.2f}. The work leading to these results has received funding from the European Union under grant agreement n°287658.

References

[1] R. Balchandran and R. Mammone, "Non-parametric estimation and correction of non-linear distortion in speech systems," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, 1998, pp. 749–752 vol.2.

[2] S. Dharanipragada and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, 2000, pp. 556–559.

[3] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *Proc.ASRU2001 - Automatic Speech Recognition and Understanding Workshop, Madonna di Campiglio*, 2001.

[4] A. De la Torre, A. Peinado, J. Segura, J. Perez-Cordoba, M. Benitez, and A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 355–366, 2005.

[5] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust speech recognition," in *in Proc. of the 7th European Conference on Speech Communication and Technology*, 2001, pp. 1135–1138.

[6] S. Furui, "Cepstral analysis technique for automatic speaker verification," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[7] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1–3, pp. 133 – 147, 1998.

[8] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Rasta-plp speech analysis technique," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 121–124.

[9] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4101–4104.

[10] V. Mitra, H. Franco, and M. Graciarena, "Damped oscillator cepstral coefficients for robust speech recognition." in *INTERSPEECH*, 2013, pp. 886–890.

[11] A. De la Torre, J. C. Segura, C. Benitez, A. Peinado, and A. Rubio, "Non-linear transformations of the feature space for robust speech recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, 2002, pp. I–401–I–404.

[12] A. J. Hudspeth, "How the ear's works work," *Nature*, vol. 341, no. 6241, pp. 397–404, 1989.

[13] M. Woszczyna, N. Aoki-Waibel, F. D. Buø, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. Rose, T. Schultz, B. Suhm, M. Tomita, and A. Waibel, "Janus 93: Towards spontaneous speech translation," in *International Conference on Acoustics, Speech, and Signal Processing 1994*, Adelaide, Australia, 1994.

[14] H. Soltau, F. Metze, C. Fugen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 214–217.