NOISE AND SPEAKER COMPENSATION IN THE LOG FILTER BANK DOMAIN

Vikas Joshi, Raghavendra Bilgi, S. Umesh

Department of Electrical Engineering
Indian Institute of Technology, Madras, India
[ee10s001, ee10s009, umeshs]@ee.iitm.ac.in

L. Garcia, C. Benitez

Dept of Signal Theory, Telematics and Communications
University of Granada, Spain
[luzgm, carmen]@ugr.es

ABSTRACT

In this paper, we propose a method to compensate for noise and speaker-variability directly in the Log filter-bank (FB) domain, so that MFCC features are robust to noise and speaker-variations. For noise-compensation, we use Vector Taylor Series (VTS) approach in the *Log FB domain*, and speaker-normalization is also done in the *Log FB domain* using Linear Vocal tract length (VTLN) matrices. For VTLN, optimal selection of warp-factor is done in Log FB domain using canonical GMM model, avoiding the two-pass approach needed by a HMM model. Further, this can be efficiently implemented using sufficient statistics obtained from the GMM and the FB-VTLN-matrices. The warp-factor selection using GMM can also be done in cepstral domain by applying DCT matrices *without* the usual approximations associated with conventional linear-VTLN. The elegance of the proposed approach is that given the speech data, we obtain directly MFCC features that are robust to noise and speaker-variations. The proposed approach, show a significant relative improvement of 31% over baseline on Aurora-4 task.

Index Terms— Speaker Normalization, Noise Compensation, VTS, TVTLN, Noise and Speaker compensation

1. INTRODUCTION

Automatic speech recognition (ASR) systems are vulnerable to both Noise and Inter-speaker variations. Several techniques for noise compensation and speaker normalization have been proposed in literature and often the efficacy of these methods are studied in isolation without considering the effect of the other. Recently, there have been some studies that attempt to compensate both noise and speaker variability and then investigate their combined effect on the recognition performance [1][2][3]. However, in most of these studies, MFCC features are first extracted from noisy speech and attempts are made to compensate for noise followed by speaker-normalization. Histogram equalization and Vector Taylor Series (VTS) are two commonly used techniques for noise-compensation, while Maximum Likelihood Linear Regression (MLLR) and VTLN are the commonly used methods used for speaker-normalization. In order to do speaker-normalization, MLLR/VTLN require an initial (first-pass) recognition which is used to estimate the normalization parameters before a final recognition is done, i.e. a two-pass approach. Recently, linear-VTLN approach has been proposed [4] which allows VTLN to be implemented as feature-transformation. However, linear-VTLN warped features are only an *approximation* to conventional-VTLN warped features since the cepstral features are truncated to usually 13 coefficients which are then used with Inverse-DCT.

This work was supported under the Indo-Spanish Joint Program of Cooperation in Science and Technology. The Indian group is supported under the projects DST/INT/SPAIN/P-5 and DST/EECE/058 of Ministry of Science and Technology. The Spanish Group is supported under project ACI2009-0892 by the Ministry of Science and Innovation.

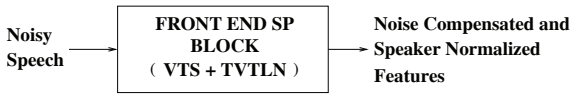


Fig. 1: Single block structure for Noise and Speaker Compensation

In this paper, we propose a method where noise and speaker-normalization are done during the feature extraction step, so that given the noisy speech data we obtain MFCC features that are noise and speaker compensated as illustrated in Fig. 1. In our proposed approach we use VTS for noise compensation and VTLN for speaker normalization with both approaches implemented in the Log FB domain. In the paper, our studies show that VTS perform better in Log FB domain compared to cepstral domain and is discussed in the section 4. In Section 2.2 we discuss the advantage of warping in Log FB domain as compared to the cepstral domain. In this approach, given Log-FB output of noisy speech, VTS returns a cleaned Log FB output. VTLN is then done by applying linear-VTLN matrix on the VTS-cleaned Log FB output to give a VTS-cleaned and VTLN-warped Log FB. Since the VTLN transformation is a square transformation, there are no truncation errors unlike linear-VTLN in cepstral domain. Finally, the two-pass approach for speaker-normalization is avoided by finding the optimal warp-factor with respect to a canonical Gaussian Mixture Model (GMM) built from VTS-cleaned Log FB coefficients. Further, the likelihood calculation for the optimal warp-factor can be efficiently implemented using sufficient statistics and the FB-warp matrices [5].

The paper is organized as follows. Section 2 briefly reviews the VTS and TVTLN approaches. In section 3 the proposed approach is presented. Section 4 has the comparison between the Log FB compensation and Cepstral domain compensation followed by experimental results and discussion in section 5. Conclusions are presented in Section 6

2. VTS AND TVTLN IN BRIEF

2.1. VTS noise compensation

The effect of additive noise on the clean speech in Log FB domain can be modelled as a non-linear transform by [6],

$$y = x + log(1 + e^{(n-x)}) = x + g(x, n) \quad (1)$$

where y is the noisy speech, x is the clean speech and n is the additive noise. In Eqn. (1), $g(x, n)$ is the non-linear function added due to presence of noise. Different variants of VTS exist depending

on the order of approximation done for $g(x,n)$. Zero order VTS (VTS-0) approximates $g(x,n)$ by a constant. VTS-1 is the first order approximation of $g(x,n)$. The MMSE estimate of clean feature $\hat{x}$, is obtained using noisy vector y, noisy speech statistics $p(y)$ and Taylor series approximated $g(x,n)$. Noisy speech statistics $p(y)$ is estimated by clean speech using the clean speech GMM $p(x)$ and noise estimate $p(n)$. The MMSE estimate is given by,

$$\hat{x} = y - \sum_{k} P(k/y) * log(1 + e^{(\hat{\mu}_n - \hat{\mu}_{xk})}) \quad (2)$$

where $P[k|y]$ is the posterior probability of the k-th Gaussian given the noisy observation, $\hat{\mu_n}$ is the estimated noise mean and $\hat{\mu_{xk}}$ is the mean of the k-th clean Gaussian. VTS can be applied both in Log FB domain as well as cepstral domain. In section 4 we study the implications of VTS in both the domains.

2.2. VTLN for Speaker Normalization

VTLN involves scaling the speech spectra to compensate for vocal tract length differences. The scaling factor is found in a maximum-likelihood (ML) framework by [7],

$$\hat{\alpha} = \underset{\alpha}{argmax}\ p(c^{\alpha}/\lambda, U) \quad (3)$$

where c^{α} is warped cepstra, λ is the reference model, U is the transcription (first pass transcription is used in test cases). Reference model λ is usually a HMM model. However, some research groups use a set of GMMs for the range of warp factors and an optimal warp factor is found by scoring unwarped MFCC with these GMMs. This avoids the 2-pass approach of HMM. Since conventional VTLN is expensive, recently a linear transformation (LT) approach (TVTLN) to VTLN, was proposed where warped features are generated by a LT [4][8], i.e.

$$c^{\alpha} = T^{\alpha} * c^{1.00} \quad (4)$$

$$T^{\alpha}_{N\times N} = D^{-1}_{N\times M} * A^{\alpha}_{M\times M} * D_{M\times N} \quad (5)$$

where $c^{1.00}$ is the unwarped cepstra, c^{α} is α warped cepstra, $T^{\alpha}_{N\times N}$ is the VTLN transformation matrix, $D^{-1}_{N\times M}$ is the rectangular Inverse DCT (IDCT) Matrix, $A^{\alpha}_{M\times M}$ is the warping matrix, N is the number of cepstral coefficients and M is the number of Log FB coefficients. Usually $N < M$ and in our experiments we use, $N = 13$ and $M = 23$. As $N < M$, the IDCT operation is an approximation. In [4], cepstral coefficients are shown to be non zero even for indices greater than 13, resulting in the errors in Log FB coefficients obtained using the IDCT operation on cepstral coefficients.

2.2.1. TVTLN in Log Filter Bank Domain

In our proposed method warping matrices are directly applied in Log FB domain avoiding the IDCT approximation errors, as explained in the Section 2.2 . Since VTS is also applied in the Log FB domain, it gives an elegant frame work to apply VTLN matrices also in Log FB domain. Thus the VTLN transformation in Log FB domain is [4]

$$[A]^{\alpha}_{k,n} = \frac{1}{2M} \sum_{l=0}^{2M-1} e^{-j\frac{2\pi}{2M}(\frac{v_l^{\alpha}}{v_s})k} e^{+j\frac{2\pi}{2M}(\frac{v_l}{v_s})n} \quad (6)$$

$$F^{\alpha}_{vts} = A^{\alpha} * F^{1.00}_{vts} \quad (7)$$

where v_l^{α} is the alpha warped frequency and v_s is the sampling frequency. Fig. 2 shows the unwarped, 0.8 warped and 1.2 warped Log FB spectra by applying transformation matrices in the Log FB domain.

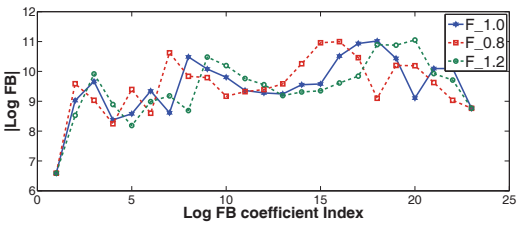


Fig. 2: Warped Log FB coefficients for $\alpha = 1.00$, $\alpha = 0.80$ and $\alpha = 1.20$ obtained by applying TVTLN in Log FB domain

3. ROBUST FEATURE GENERATION USING VTS AND TVTLN IN LOG FB DOMAIN

Feature Extraction : Fig. 1 describes the robust feature generation which involves 2 steps. First, VTS compensated Log FB features are obtained. Then, VTLN is performed on VTS-cleaned Log FB features as explained below:

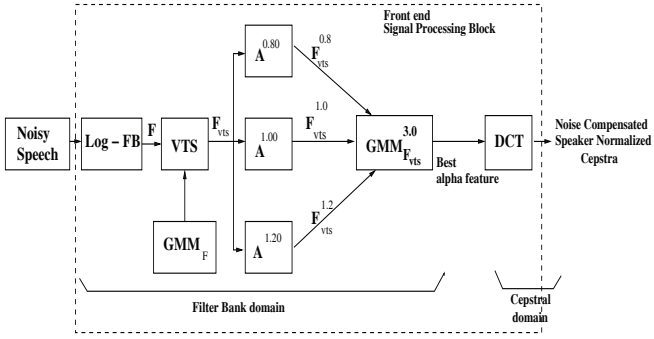


Fig. 3: VTS+TVTLN (GMM-FB) feature generation approach. Warp factor estimation is done using GMM in Log FB domain

- **VTS compensation :** VTS compensation needs clean speech GMM as mentioned in section 2.1. GMM for VTS compensation (GMM_F) is built using clean speech Log FB features (F). GMM_F is built in the training phase. Then, given the noisy speech signal, VTS compensated Log FB features are obtained (F_{vts}) using GMM_F as shown in Fig. 3 and Eqn. (2).
- **VTLN Warping :** Warping is done in Log FB domain by multiplying the VTS compensated Log FB coefficients (F_{vts}) with warping matrices ($A^{0.8}$) to get VTS-compensated and VTLN warped features (F^{α}_{vts}). Since the warping matrices are applied in Log FB domain, *IDCT approximation errors are eliminated*. The best VTLN warped feature is estimated in ML sense according Eqn. (3). The reference model used for best warp factor estimation is the GMM model built using VTS-compensated, unwarped features (F_{vts}). This GMM model is built iteratively. In the first iteration unwarped features are used to build $GMM^{1.0}_{F_{vts}}$. In the subsequent iterations previous iteration best warp features are used to build a canonical GMM model. $GMM^{3.0}_{F_{vts}}$ shown in the Fig. 3 refer to the third iteration GMM model. This process of building GMM model is done only in the training phase. Thus best warp features are estimated from 21 warped features and $GMM^{3.0}_{F_{vts}}$ to get VTS-compensated, best VTLN warped Log FB feature. This can be efficiently implemented using sufficient statistics and warp matrices.

- Finally DCT is applied on VTS-cleaned, best alpha warped Log FB features to obtain noise compensated and speaker normalized cepstra.

HMM model is trained using the above VTS compensated, best VTLN warped features. Testing phase is simplified by using the above feature generation technique. During test, from the noisy speech signal, noise compensated and speaker normalized features are extracted and the recognition is done using the HMM model built in the training phase. This approach of warp factor estimation in the Log FB domain is referred to as VTS+TVTLN (GMM-FB). This approach is more elegant since VTS compensation, VTLN warping and best warp estimation are all done in Log FB domain itself.

3.1. Warp factor estimation in Cepstral domain using VTS and TVTLN in the Log FB domain

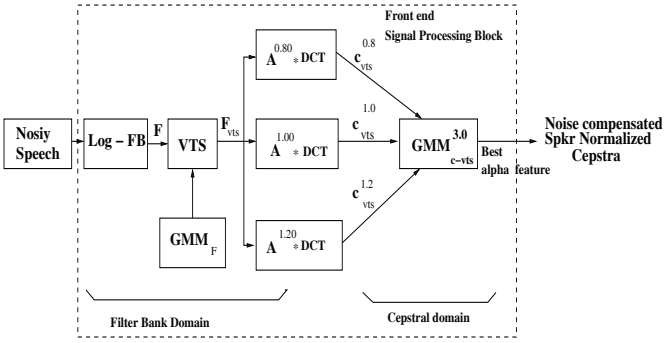


Fig. 4: VTS+TVTLN (GMM-CEP) feature generation approach. Warp factor estimation is done using GMM in cepstral domain.

In this section we investigate the efficacy of warp factor estimation after the application of DCT. The advantage of performing the warp factor estimation in cepstral over Log FB domain is that, the cepstral features are fairly uncorrelated and can be modelled better using diagonal covariance GMMs. Warp factor estimation in cepstral domain is easily done by multiplying the pre-computed warping matrices with DCT matrix as shown in the Fig. 4. Even in this approach *VTS and VTLN warping are done in Log FB domain itself and hence will not have IDCT approximation errors.* Only the warp factor estimation is done in the cepstral domain. In this approach warped cepstra are directly obtained from VTS compensated Log FB features by,

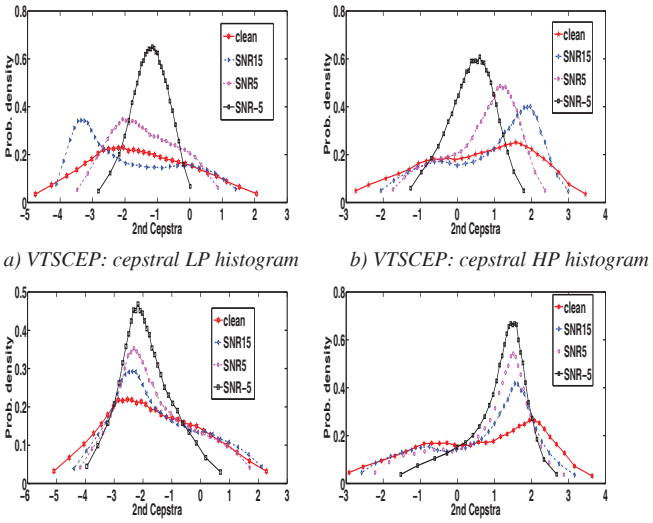
$$c^{\alpha}_{vts} = D * [A^{\alpha} * F^{1.00}_{vts}] \quad (8)$$

where c^{α}_{vts} is the VTS compensated, VTLN warped feature. This method is referred to as VTS+TVTLN (GMM-CEP). Feature generation steps are same as explained for VTS+TVTLN (GMM-CEP), except that the best alpha estimation is done directly in the cepstra domain.

3.1.1. Summary of the proposed approach

Our proposed approach of feature extraction have following merits :

1. It is fast and more convenient. Both noise compensation and speaker normalization are done in the feature domain.
2. Use of GMM in place of HMM makes VTLN faster and simple (fast likelihood calculation, 1 pass approach).
3. VTS applied in Log FB domain performs better than VTS in cepstral domain (Section 4).
4. TVTLN in Log FB domain eliminates the IDCT approximation errors.



a) VTSCEP: cepstral LP histogram *b) VTSCEP: cepstral HP histogram*

c) VTSFB: cepstral LP histogram *d) VTSFB: cepstral HP histogram*

Fig. 5: Histograms of 2nd cepstral coefficient for a) VTSCEP: cepstral LP histogram, b) VTSCEP: cepstral HP histogram, c) VTSFB: cepstral LP histogram, d) VTSFB: cepstral HP histogram

4. VTS IN FILTER BANK DOMAIN VERSUS CEPSTRAL DOMAIN

VTS can be applied both in cepstral domain (VTSCEP) and Log FB domain (VTSFB). VTS model for noisy cepstra is given by,

$$c_y = c_x + D * log(1 + e^{D^{-1}(c_n - c_x)}) \quad (9)$$

where c_y is the noisy cepstra, c_x is the clean cepstra, c_n is the noise vector due to additive noise at the input and D is the DCT matrix. The motivation for the Log FB compensation is that different frequency bands have different SNR levels which makes compensation of individual filter bank energies more appropriate. In this section, we examine the noise compensating capabilities of VTS in Log FB domain and cepstral domain by analysing the histogram plots for cepstral coefficients obtained from both the approaches. The basic idea of this analysis, is that for any good "noise compensating" method, histogram of "cleaned" features should match those of the original speech features. In [9], we show the importance of histogram matching of LP and HP cepstra, wherein a significant improvement in the recognition accuracy was obtained by equalizing the LP and HP histograms.

Here, noise compensated cepstral features are obtained for VTSFB and VTSCEP approaches. The histogram plots, as shown in Fig. 5 are obtained for low pass (LP) filtered and high pass (HP) filtered cepstral coefficients at different SNR levels. Filtering is done in cepstral domain *for each frame* by a simple averaging and differencing operation. More detailed explanation is present in [9].

$$c_{lp}(n) = [c(n) + c(n-1)]/2 \quad n = 1, 2, ..12 \quad (10)$$

$$c_{hp}(n) = [c(n) - c(n-1)]/2 \quad n = 1, 2, ..12 \quad (11)$$

where $c(n)$ is the n^{th} cepstral coefficient *of a particular frame*, c_{lp} is LP filtered and c_{hp} is HP filtered part of $c(n)$. Here we compare the LP and HP cepstral histograms for VTSFB and VTSCEP approaches. For VTSCEP, there is a considerable mismatch between the clean histogram and noisy histogram for both LP and HP cepstral coefficients. However, for VTSFB the means of the histogram are well compensated even for SNR-5dB. Thus VTSFB features seem to

Table 1: Recognition Results on Aurora 4 database

Test Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Avg	R.I%
Baseline	87.61	75.42	53.30	53.17	46.95	56.57	45.4	76.89	64.21	45.28	41.98	36.26	47.51	36.45	54.9	0
VTS	88.21	83.28	68.86	62.62	62.34	67.35	62.79	82.65	77.28	62.32	55.39	55.28	61.83	59.16	67.81	23.5
VTS+TVTLN (HMM)	90.57	86.42	73.08	67.83	66.39	72.43	68.78	86.31	81.82	68.56	60.21	61.57	66.86	63.85	72.48	32.02
VTS+TVTLN (GMM-FB)	89.74	85.3	72.67	66.36	63.96	71.68	65.72	84.74	79.58	65.37	58.08	56.9	64.26	61.07	70.39	28.22
VTS+TVTLN (GMM-CEP)	90.64	86.19	73.32	67.36	65.83	72.71	67.01	85.80	81.09	69.06	58.38	59.41	66.65	63.48	71.92	31

Table 2: Recognition Results averaged over set A, set B and set C for Aurora 2 database

	BaseLine	VTSCEP	VTSFB	VTS+TVTLN (HMM)	VTS+TVTLN (GMM-FB)	VTS+TVTLN (GMM-CEP)
clean	99.31	99.20	99.25	99.29	99.26	99.27
20 dB	97.92	96.38	98.26	98.56	98.23	98.47
15 dB	94.59	94.53	96.85	97.19	96.74	97.03
10 dB	83.06	90.74	92.84	93.35	92.56	92.97
5 dB	54.34	77.60	81.55	81.94	80.95	81.09
0 dB	25.13	49.26	54.13	53.75	53.35	53.31
-5 dB	12.91	20.32	22.89	21.74	22.75	22.78
Avg SNR $>$ 5dB	93.72	95.25	96.8	97.1	96.7	96.95
Overall Avg	71.01	81.70	84.73	84.96	84.37	84.57

be better compensated compared to VTSCEP, which is also reflected in the recognition results shown in table 2 for Aurora-2 databases.

5. EXPERIMENTAL RESULTS

5.1. Experimental Setup

The proposed method is tested on AURORA-2 and AURORA-4 databases. Feature generation, training and testing procedures are as explained in the section 3. In AURORA2 connected digits task, each digit is modelled as a left to right continuous density HMM with 16 states and 6 Gaussians per state. 13 dimensional MFCC feature is used as the basic parametrization of the speech signal using C_0 instead of the logarithmic energy. First and second order regressions are augmented to 13 MFCC vectors, yielding a final 39 component feature vector. CMS is performed by sentence-by-sentence subtraction of the mean values of each cepstral coefficient. For Aurora-4, recognition system is based on continuous cross-word triphone models with 3 tied states and a mixture of 6 Gaussians per state. The language model is the standard bi-gram for the WSJ0 task. GMM for VTS compensation is obtained from clean speech Log FB coefficients. GMM for warp factor estimation is obtained from clean speech Log FB coefficients after applying VTS. In all our experiments VTS with zero order approximation (VTS-0) is used. Although VTS-0 performs inferior to VTS-1, it is computationally efficient.

5.2. Discussion

Tables 1 and 2 show the results for the Aurora-4 and Aurora-2 databases. VTSFB refers to the VTS applied in the Log FB and VTSCEP for VTS in the cepstral domain. In rest of experiments only VTS imply VTSFB. VTS+TVTLN (HMM) is combination of VTS and TVTLN with the warp factor estimation is done using traditional 2 pass HMM approach. In VTS+TVTLN (GMM-FB) warp factors are estimated in Log FB domain itself. VTS+TVTLN (GMM-CEP) does the warp factor estimation in the cepstral domain. Recognition results show a very small degradation in performance due to use of GMM in place of HMM for warp factor estimation, while gaining significantly in terms of computational efficiency. Performance of VTS+TVTLN(GMM-FB) is slightly inferior compared to HMM and GMM-CEP approach, due to the higher correlation in the Log FB features. Note that GMM-FB and GMM are very similar in implementation, with the only difference being the estimation of warp factor before or after DCT. Results show a significant improvement in recognition accuracy for all 3 approaches (HMM, GMM-FB and GMM-CEP) for Aurora-4 database. For Aurora-2 database, VTS+TVTLN (HMM) and VTS+TVTLN (GMM-CEP) perform well for SNR $>$ 5dB.

6. CONCLUSIONS

In the proposed approach, both noise and speaker compensation are done in the Log FB domain. The advantage of this approach is that, noise and speaker robust MFCC features are derived from noisy speech signal directly during the feature extraction. VTLN warp factor estimation is made computationally efficient by using GMM based warp factor estimation and sufficient statistics base likelihood calculation. GMM-CEP based VTS+TVTLN approach is computationally efficient and shows a relative improvement of 12.8% and 4.7% in WER over VTS alone for Aurora-2 and Aurora-4 databases respectively, making it well suited for real time applications.

7. REFERENCES

[1] Vikas Joshi, R. Bilgi, S. Umesh, L. Garcia, and C. Benitez, "Efficient speaker and noise normalization for robust speech recognition," in *Proc. Interspeech*, 2011.

[2] L Garcia, C Benitez, J C Segura, and S Umesh, "Combining speaker and noise feature normalization techniques for automatic speech recognition," in *Proc. of ICASSP-2011*, May 2011.

[3] K.K. Chin, H. Xu, M. J. F. Gales, C. Breslin, and K. Knill, "Rapid joint speaker and noise compensation for robust speech recognition," in *Proc. ICCASP*, 2011.

[4] S. Umesh, András Zolnay, and Hermann Ney, "Implementing frequency-warping and vtln through linear transformation of conventional mfcc," in *Interspeech*, 2005, pp. 269–272.

[5] P. T. Akhil, S. P. Rath, S. Umesh, and D. R. Sanand, "A computationally efficient approach to warp factor estimation in vtln using em algorithm and sufficient statistics," in *Proc. Interspeech*, 2008.

[6] Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. ICASSP-96*, 1996, pp. 733–736.

[7] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Process.*, , no. 6, pp. 49–60, 1998.

[8] S. Panchapagesan and Abeer Alwan, "Frequency warping for vtln and speaker adaptation by linear transformation of standard mfcc," *Comput. Speech Lang.*, vol. 23, pp. 42–64, January 2009.

[9] Vikas Joshi, R. Bilgi, S. Umesh, L. Garcia, and C. Benitez, "Sub band level histogram equalization for robust speech recognition," in *Interspeech*, 2011.