SPECTRO-TEMPORAL FEATURES FOR NOISE-ROBUST SPEECH RECOGNITION USING POWER-LAW NONLINEARITY AND POWER-BIAS SUBTRACTION

Shuo-Yiin Chang[1,2], *Bernd T. Meyer*[1,3], *Nelson Morgan*[1,2]

[1]International Computer Science Institute, Berkeley, CA, USA
[2]EECS Department, University of California-Berkeley, Berkeley, CA, USA
[3]Medical Physics, University of Oldenburg, Germany

shuoyiin@eecs.berkeley.edu, bmeyer@icsi.berkeley.edu, morgan@icsi.berkeley.edu

ABSTRACT

Previous work has demonstrated that spectro-temporal Gabor features reduced word error rates for automatic speech recognition under noisy conditions. However, the features based on mel spectra were easily corrupted in the presence of noise or channel distortion. We have exploited an algorithm for power normalized cepstral coefficients (PNCCs) to generate a more robust spectro-temporal representation. We refer to it as power normalized spectrum (PNS), and to the corresponding output processed by Gabor filters and MLP nonlinear weighting as PNS-Gabor. We show that the proposed feature outperforms state-of-the-art noise-robust features, ETSI-AFE and PNCC for both Aurora2 and a noisy version of the Wall Street Jounal (WSJ) corpus. A comparison of the individual processing steps of mel spectra and PNS shows that power bias subtraction is the most important aspect of PNS-Gabor features to provide an improvement over Mel-Gabor features. The result indicates that Gabor processing compensates the limitation of PNCC for channels with frequency-shift characteristic. Overall, PNS-Gabor features decrease the word error rate by 32% relative to MFCC and 13% relative to PNCC in Aurora2. For noisy WSJ, they decrease the word error rate by 30.9% relative to MFCC and 24.7% relative to PNCC.

Index Terms— spectro-temporal features, robust speech recognition, large vocabulary speech recognizion

1. INTRODUCTION AND RELATED WORK

Although state-of-the-art automatic speech recognition (ASR) systems can achieve high performance in clean environments, performance degrades in the presence of noise. Several existing methods focus on compensating the difference between clean speech and noisy speech in different aspects such as model-based[1] [1][2] or feature-based approaches [3][4][5]. Unlike ASR systems, human listeners rely on attention-driven (cognitive) selection of a specific speaker, e.g., in a high-noise cocktail party situation, which results in high recognition scores for human listeners, and which inspires researchers to find more robust features based on biological models of the auditory system. Over the last decade, a number of physiological experiments on different mammalian species have revealed that neurons in the primary auditory cortex are sensitive to particular spectro-temporal patterns referred to as spectro-temporal receptive fields (STRFs) [6]. Based on this evidence, spectro-temporal features, which serve as a model for STRFs, have been applied to ASR. Several studies have successfully incorporated Gabor function approximations into ASR [7][8][9]. In general, these approaches define a series of spectral, temporal, and spectro-temporal modulation filters that can be seen as very roughly modeling neuron firing patterns for particular spectro-temporal signal components. Purely temporal features such as TRAPS [10] and HATS [11] can be regarded as special cases of spectro-temporal features. Gabor filters have also been used for speech and nonspeech discrimination [12][13]. While a number of experiments have shown the efficiency of using spectro-temporal filters for ASR, mel spectra may not be the optimal time-frequency representation to be filtered for this purpose.

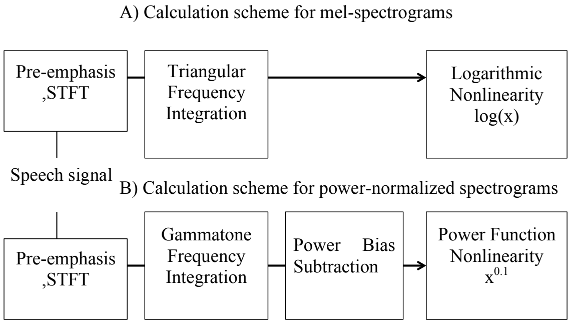
A) Calculation scheme for mel-spectrograms
Pre-emphasis ,STFT
Triangular Frequency Integration
Logarithmic Nonlinearity log(x)
Speech signal
B) Calculation scheme for power-normalized spectrograms
Pre-emphasis ,STFT
Gammatone Frequency Integration
Power Bias Subtraction
Power Function Nonlinearity $x^{0.1}$

Figure 1: Comparsion of mel spectrum and PN spectrum

Shao [14] suggests that a gammatone filter bank based on the auditory periphery model is robust for noisy speech recognition. In [4], Kim and Stern propose the power normalized cepstral coefficient (PNCC) algorithm using gammatone filters followed by power bias subtraction and power nonlinearity compression. PNCC is relatively insensitive to stationary noise. Therefore, spectra generated by the PNCC algorithm could be a good choice as the representation to be filtered. We refer to the spectra generated from PNCC as the power normalized spectrum (PNS). Here we present a series of ASR experiments comparing mel spectra to power normalized spectra when processed with Gabor filters.

[1] Note that model-based approaches can be applied in addition to the feature-oriented methods described here or elsewhere.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

Due to the use of a large number of Gabor filters, our feature dimension is very high, which can create difficulties in current systems. Marki and Stylianous [15] employ a variety methods for dimensionality reduction. Zhao and Morgan [16] divide features into several streams so that each represents a patch of information in the spectro-temporal field. In this paper, we employ tandem MLP-HMM acoustic models [17] to integrate the features for speech recognition. The PNS filtered by Gabor filters and MLP processing are referred to as PNS-Gabor features.

The experiments presented in this paper show that PNS-Gabor plus MFCC are more robust than ETSI-AFE, PNCC and Mel-Gabor plus MFCC for both Aurora2 and a noisy version of WSJ. Further, we perform an analysis of the importance of individual signal processing steps differentiating mel spectra and power normalized spectra that result in the increased robustness.

2. METHODS

The PNS-Gabor features we propose in this paper comprise three steps: (1) Calculate the power normalized spectrum (2) Process the spectra with a Gabor filter bank (3) Apply non-linear processing with MLPs. We describe the three parts in the following sections.

2.1 Power normalized spectrum

PNCC differs from MFCC in three respects (Fig. 1): (1) gammatone filter (2) medium-duration bias subtraction (3) power-law nonlinearity. First, PNCC employs gammatone auditory filters based on equivalent rectangular bandwidth. Gammatone filters are derived from psychophysical observations of the auditory periphery, i.e., the filter bank represents a model of cochlear filtering. We use 30 channels for 8 kHz corpus and 40 channels for 16 kHz corpus as suggested by [4]. For mel spectra, we use 23 channels defined in the ETSI standard for MFCC calculation [18]. Second, a subtraction of the medium-duration power bias is carried out, where the bias level calculation was based on the ratio of arithmetic mean and geometric mean (AM-GM ratio) of the medium duration power, which is motivated by a decrease of the AM-GM ratio for reduced noise power. Finally, a power nonlinearity with an exponent of 0.1 replaces the logarithm nonlinearity for compression. This nonlinearity might be a better model for threshold effects of auditory fire rate responses. According to the observation in [19], the auditory nerve firing rate is constant when the input sound pressure level is below -10 dB, while the output of the logarithm would be dominated by noise when the intensity of the input signal is low. An example of mel spectrum and power normalized spectrum is shown in Fig. 2 illustrating greater insensitivity to noise for the power normalized spectrogram. In the following experiments, the effect of these differences between MFCC and PNCC features is independently analyzed.

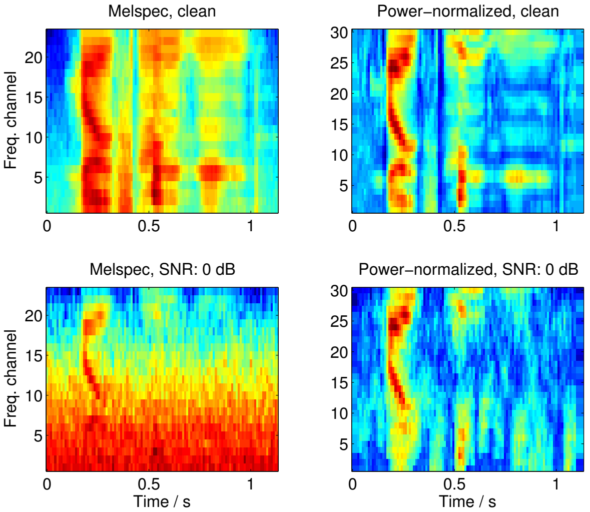

Figure 2 : Clean and noisy mel spectrogram (left) and power normalized spectrogram (right).

2.2 Gabor features

PNS Gabor features are obtained by convolving two dimensional modulation filters and the PNS. To generate filters serving as model for spectro-temporal receptive fields (STRFs), we multiply a complex sinusoid with a Hanning envelope. The complex sinusoid (with time modulation frequency ω_n and spectral modulation frequency ω_k) is represented as:

$$s(n,k) = \exp[i\omega_n(n-n_0) + i\omega_k(k-k_0)] \qquad (1)$$

while the Hanning envelope is given (with W_n and W_k denote window length)

$$h(n,k) = 0.5 - 0.5 \cdot \cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right) \cdot \cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right) \qquad (2)$$

By tuning parameters of spectral and temporal modulation frequency, Gabor functions have different extent and orientation for a given number of oscillations under the envelope as used in this study. In particular, if the spectral modulation frequency is set to zero, only temporal modulation filtering is performed. The Gabor filter bank used here has been adapted from [20]. The 59 Gabor filters with the corresponding temporal and spectral modulation frequencies are shown in Fig. 3. Filters with a large spectral extent result in high correlations between frequency channels (i.e., moving large filters by only one channels has only small influence on the filter output). Hence, a subset of the possible combinations are used to avoid high correlations of feature components, resulting in 631-dimensional vectors using 30 channels of the power normalized spectrum and 814-dimensional vectors using 40 channels of power normalized spectrum. (A 449-dimensional vector is used for the case of Gabor-filtered mel spectrum.)

2.3 MLP processing

MLPs can handle correlated and high-dimensional features with few distributional assumptions. As is typical in tandem processing, an MLP is trained to generate posterior phone probabilities, followed by computing the logarithm and PCA to yield features used as input to HMMs. Means and variances were normalized per utterance before HMM training and testing.

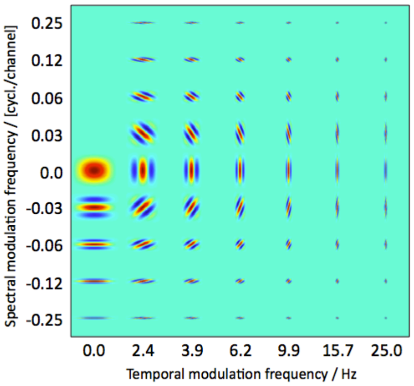

Figure 3: 2D-filters to the PN spectrum, shown by temporal and spectral modulation frequencies.

3. EXPERIMENTAL SETUP

The approach proposed here is evaluated with the Aurora2 testing environment covering the recognition of noisy digits, and with a noisy version of the Wall Street Journal (WSJ) corpus. For Aurora2, we use the clean connected digits for training. Three testing sets (set A, B and C) are used with clean and noisy data. The testing data set A covers four different noise types (subway, babble, exhibition and car), while the testing data B covers four different noise types (restaurant, street, airport and train station). The testing set C covers two noise types respectively from set A and set B (subway and street), plus additional convolution noise. Different SNR values ranging from 0 dB to 20 dB were tested in each case. The HMMs were configured according to [21]: whole-word HMMs with 16 states and with 3-Gaussian mixture with diagonal covariance per state. Baseline results are obtained with the standard Aurora2 MFCC frontend (WI007) [18], which converted each signal frame into 13 cepstral coefficients, with subsequent addition of first and second derivative and utterance-wise mean and variance normalization. The average word error rates (WERs) of this task are obtained by averaging over WERs of the test sets.

For WSJ, we started out with clean data taken from a 77.8 hour WSJ1 dataset (284 speakers) for training and an 0.8 hour WSJ-eval94 dataset (20 speakers) for testing. Estimated additive and channel noise from real-world recordings was applied to both training and testing dataset using the "renoiser" tool [22]. Designed for use in the DARPA RATS project, the system analyzes data from RATS rebroadcast example signals (in this case, LDC2011E20) to estimate the noise characteristic including SNRs and frequency-shifts; the original data is described in [23] and consists of a variety of continuous speech sources that have been transmitted and received over 8 different radio channels, resulting in significant signal degradations. The 8 different radio channel characteristics were estimated by the renoiser tool as specified in Table 3. We applied the same noise characteristics to WSJ data yielding what we refer to as "re-noised WSJ".

We use the HTK toolkit [24] for both training and decoding with both data bases. The acoustic models are cross-word triphones estimated with maximum likelihood. Except for silence, each triphone is modeled using a three-state HMM. The resulting triphone states are clustered using decision trees to 5000 tied states, each of which was modeled by 32-component of Gaussian mixture model. We use version 0.6 of the CMU pronunciation dictionary (stress removed) and the standard 5k bigram language model created at Lincoln Labs for the 1992 evaluation.

All MLPs were trained with a temporal context window of 9 successive frames. We used 160 hidden nodes for Aurora2 and 500 hidden nodes for re-noised WSJ. In Aurora2, the output layer consisted of 56 context-independent phonetic targets while 41 context-independent phonetic targets were defined for noisy WSJ. MFCCs $+\Delta$ $+\Delta\Delta$ were then concatenated with Gabor features. The dimension of Gabor features is then reduced via PCA to 32, resulting in a 71-dimension feature vector.

4. RESULTS AND DISCUSSION

In Table 1, we compare several different configurations of PNS-Gabor features and Mel-Gabor features after concatenating MFCC in Aurora2. The result for Mel-Gabor features plus MFCC is presented in row (2), which is 15% relative better than the MFCC baseline. From row (3) to row (6), the results were obtained from deconstructing PNS into four different configurations. In row (7), instead of being filtered by Gabor filters, PNCC was used as input for MLP, from which we could investigate the benefit of combining MFCC and PNCC without Gabor filtering. As shown in row (3), we only switched from mel filter banks to gammatone filter banks. We referred to it as GT(*l*)-Gabor. In row (4), gammatone filter banks were further processed by power nonlinearity compression *p* instead of logarithm compression *l*, which was referred as GT(*p*)-Gabor. PNS(*l*)-Gabor features were obtained by performing a power bias subtraction followed by logarithmic compression. The result for PNS-Gabor features is presented in row (6). GT(*l*)-Gabor didn't perform as well as conventional Mel-Gabor features, while GT(*p*)-Gabor gave a slight improvement. This implies that the power nonlinearity is helpful to inhibit the effects of noise, as expected.

	Filter bank	Pow Sub	Com.	Gb.	WER
(1) MFCC	-	-	-	-	18.14
(2) Mel-Gb + MFCC	Mel	no	log	yes	15.41
(3) GT(*l*)-Gb + MFCC	GT	no	log	yes	15.75
(4) GT(*p*)-Gb + MFCC	GT	no	pow	yes	14.86
(5) PNS(*l*)-Gb + MFCC	GT	yes	log	yes	13.12
(6) PNS-Gb + MFCC	GT	yes	pow	yes	**12.30**
(7) PNCC-MLP + MFCC	GT	yes	pow	no	14.06

Table 1: Aurora2 WER of Gabor features with different spectro-temporal representations. The baseline is a 39-dimensional MFCC plus the first 2 derivatives with mean and variance normalization.

However, as shown in Table 1, the most effective step is power bias subtraction, from which we got 16.7% relative improvement by comparing GT(*l*)- Gabor and PNS(*l*)-Gabor features. The best result came from the PNS-Gabor feature, which is 20% relatively better than Mel-Gabor feature. Even after power bias subtraction, power nonlinear compression can help. In row (7), we showed the WER from combination of PNCC and MFCC using MLP, which is significantly worse than the proposed Gabor-filtered PNCC augmentation of MFCCs; the latter is 15.7% better.

In Table 2, the proposed front end is compared with other noise robust features, ETSI-AFE [5] and PNCC [4]. WER of Mel-Gabor features and PNS-Gabor features plus MFCC are presented in row (4) and row (5). The best result was obtained by concatenating MFCC and PNS-Gabor features, which is 13.3% better than PNCC and 6.0% better than ETSI-AFE. Fig. 4 and Fig. 5 provide a more detailed comparison by different SNRs and noise type respectively. As shown in Fig. 4, the improvements of PNS-Gabor plus MFCC was much more evident for noisier conditions. From Fig. 5, PNS-Gabor plus MFCC performed better for all the environments.

	WER	Rel to (1)	Rel to (2)
(1) MFCC	18.14	-	-27.8%
(2) PNCC	14.19	21.7%	-
(3) ETSI-AFE	13.09	27.8%	7.8%
(4) Mel-Gb + MFCC	15.41	15%	-8.6%
(5) PNS-Gb + MFCC	**12.30**	32.2%	13.3%

Table 2: WER of noise-robust features for Auroa2

Beyond the experiments conduced on the small vocabulary Aurora2 task, we compared the features in re-noised WSJ for multi-conditioned training and testing. The WER is shown in Table 4. On average, Gabor features (Mel-Gabor plus MFCC and PNS-Gabor plus MFCC) are significantly better than other features for most channels, especially for channels with a frequency shift characteristic (channel D and channel H). PNS-Gabor plus MFCC is relative 30.9% better than MFCC and 24.7% better than PNCC.

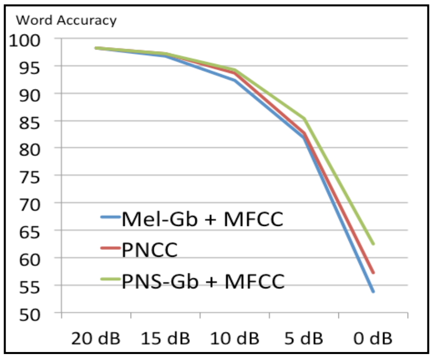

Figure 4: Average word accuracy of PNS-Gabor plus MFCC, Mel-Gabor plus MFCC and PNCC under different SNR conditions for the Aurora 2 task.

5. CONCLUSION

In this paper, we employed a more robust spectro-temporal representation incorporating key parts of the PNCC algorithm, augmented by Gabor filtering and an MLP. These PNS-Gabor features improved WER for both small and large vocabulary noisy recognition tasks. It appears that power bias subtraction and Gabor filtering are the key steps for decreasing the WER. PNS-Gabor features were particularly effective for frequency-shifted channels in the larger task. Extending the feature vector to include PNS-Gabor gave around 30% improvement relative to the MFCC baseline and yielded WER 13-25% better than PNCC.

6. ACKNOWLEDGEMENTS

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. Bernd T. Meyer's work is supported by a post-doctoral fellowship of the German Academic Exchange Service (DAAD). We thank Chanwoo Kim and Richard Stern for use of PNCC. We thank Dan Ellis (Columbia University and the International Computer Science Institute [ICSI]) for the renoiser tool.

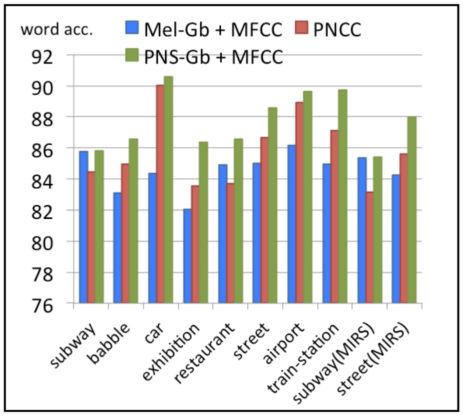

Figure 5: Average word accuracy of PNS-Gabor plus MFCC, Mel-Gabor plus MFCC and PNCC under different environments for the Aurora 2 task.

	Microphone	SNR	Frequency shift
Channel A	Motorola HT1250	15.6	0
Channel B	Midland GXT1050	6.2	0
Channel C	Midland GXT1050	6.0	0
Channel D	Galaxy DX2547	3.5	180.9 Hz
Channel E	Icom IC-F70D	0.9	0
Channel F	Trisquare TSX300	3.0	0
Channel G	Vostek LX-3000	18.7	0
Channel H	Magnum 1012 HT	3.0	120.7 Hz

Table 3: channel characteristic in re-noised WSJ

WER	MFCC	ETSI-AFE	PNCC	Mel-Gb +MFCC	PNS-Gb +MFCC
Clean	19.06	19.22	16.21	16.87	**14.04**
Channel A	42.22	36.89	32.60	30.53	**26.78**
Channel B	49.25	48.44	46.79	41.66	**35.09**
Channel C	51.05	49.38	47.15	41.72	**34.79**
Channel D	59.52	54.49	58.18	46.48	**40.35**
Channel E	78.43	74.05	72.93	63.51	**55.46**
Channel F	65.58	62.48	60.41	51.23	**44.28**
Channel G	23.55	23.25	20.72	21.56	**17.24**
Channel H	56.23	51.56	53.01	44.86	**39.39**
Average	49.4	46.6	45.3	39.8	**34.2**

Table 4: WER of noise-robust features for re-noised WSJ

7. REFERENCES

[1] O. Kalinli, M.L. Seltzer, and A. Acero, “Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition,” in Proc. ICASSP, 2009, pp. 3825-3828

[2] F. Flego and M. J. F. Gales, “Factor Analysis Based VTS Discriminative Adaptive Training” Proc. ICASSP. IEEE, 2012, pp. 4669–4672.

[3] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, “Normalized amplitude modulation features for large vocabulary noise-robust speech recognition”, Proc. ICASSP 2012, March 2012, pp. 4117-4120

[4] C. Kim and R. M. Stern, “Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring”, in Proc. ICASSP, pp. 4574–4577, 2010.

[5] Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Adv. Front-end Feature Extraction Algorithm; Compression Algorithms, ETSI ES 202 050 Ver. 1.1.5, 2007

[6] N. Mesgarani, and S. Shamma, “Speech Processing with a Cortical Representation of Audio”, Proc. ICASSP 2011, May 2011, pp. 5872-5875.

[7] M. Kleinschmidt, “Localized spectro-temporal features for automatic speech recognition,” in Proc. of Eurospeech, 2003, Sep 2003, pp. 2573–2576.

[8] F. Joublin X. Domont, M.Heckmann and C. Goerick, “Hierarchical spectro-temporal features for robust speech recognition” Proc. ICASSP 2008, March 2008, pp. 4417-4420

[9] S. Ravuri and N. Morgan, “Using spectro-temporal fea- tures to improve AFE feature extraction for automatic speech recognition,” in Proc. ICASSP, 2010, March 2010, pp. 1181–1184.

[10] H. Hermansky and S. Sharma, “Temporal patterns (TRAPs) in ASR of noisy speech,” Proc. ICASSP 1999, March 1999, pp. 289-292 vol. 1.

[11] B.Y. Chen, Q. Zhu, and N. Morgan, “A Neural Network for Learning Long-Term Temporal Features for Speech Recognition,” Proc. ICASSP 2005, March 2005, pp. 945-948

[12] N. Mesgarani, M. Slaney, and S. A. Shamma, “Discrimination of Speech From Nonspeech Based on Multiscale Spectro-Temporal Modulations”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 3, May 2006.

[13] T. J. Tsai and N. Morgan “ Longer Features: They Do a Speech Detector Good” Proc. Interspeech 2012

[14] Y. Shao, Z. Jin, D. Wang and S. Srinivasan “An Auditory-based Feature for Robust Speech Recognition” Proc. Interspeech 2009, Sep 2009, pp. 4625-4628.

[15] M. Marki and Y. Stylianou, “Discrimination of speech from nonspeech in broadcast news based on modulation frequency features,” Speech Communication 53, pp. 726-735, 2011.

[16] S. Ravuri S. Zhao and N. Morgan, “Multi-stream to many-stream: Using spectro-temporal features for automatic speech recognition,” Proc. of Interspeech, 2009, Sep 2009, pp. 2951-2954.

[17] H. Hermansky, D. Ellis, and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems”, Proc. ICASSP, Istanbul, Turkey, 2000, June 2000, pp. 1635-1638 vol.3.

[18] ETSI Standard, “Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms”, ETSI ES 201 108 v.1.1.2, Apr. 2000

[19] X. Zhang, M. G. Heinz, I. C. Bruce, and, L. H. Carney, “A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression,” J. Acoust. Soc. Am., vol. 109, no. 2, pp. 648–670, Feb 2001.

[20] B. Meyer, C. Spille, B. Kollmeier and N. Morgan, “Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition”, in Proc. Interspeech 2012

[21] H. Hirsch, and D. Pearce. “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” Proc. of ICSLP, volume 4, pp.29-33

[22] D. Ellis, http://labrosa.ee.columbia.edu/projects/renoiser/create_wsj.html

[23] K.Walker and S. Strassel, “The rats radio traffic collection system,” in Proc. of ISCA Odyssey, 2012.

[24] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK Book. Cambridge,United Kingdom: Entropic Ltd.