Robust Automatic Speech Recognition Using PD-MEEMLIN

Igmar Hernández[1],Paola García[1],Juan Nolazco[1],Luis Buera[2],Eduardo Lleida[2]

[1] Computer Science Department, Tecnolgico de Monterrey, Campus Monterrey, México.
[2] Communications Technology Group (GTC), I3A, University of Zaragoza, Spain.
{A00778595, paola.garcia, jnolazco,}@itesm.mx and {lbuera,lleida}@unizar.es

Abstract. This work presents a robust normalization technique by cascading a speech enhancement method followed by a feature vector normalization algorithm. To provide speech enhancement the Spectral Subtraction (SS) algorithm is used; this method reduces the effect of additive noise by performing a subtraction of the noise spectrum estimate over the complete speech spectrum. On the other hand, an empirical feature vector normalization technique known as PD-MEMLIN (Phoneme-Dependent Multi-Enviroment Models based LInear Normalization) has also shown to be effective. PD-MEMLIN models clean and noisy spaces employing Gaussian Mixture Models (GMMs), and estimates a set of linear compensation transformations to be used to clean the signal. The proper integration of both approaches is studied and the final design, PD-MEEMLIN (Phoneme-Dependent Multi-Enviroment Enhanced Models based LInear Normalization), confirms and improves the effectiveness of both approaches. The results obtained show that in very high degraded speech PD-MEEMLIN outperforms the SS by a range between 11.4% and 34.5%, and for PD-MEMLIN by a range between 11.7% and 24.84%. Furthemore, in moderate SNR, i.e. 15 or 20 dB, PD-MEEMLIN is as good as PD-MEMLIN and SS techniques.

1 Introduction

The robust speech recognition field plays a key rule in real environment applications. Noise can degrade speech signals causing nocive effects in Automatic Speech Recognition (ASR) tasks. Even though there have been great advances in the area, robustness still remains an issue. Noticing this problem, several techniques have been developed over the years, for instance the Spectral Subtraction algorithm (SS) [1]; and in the last decade, SPLICE (State Based Piecewise Linear Compensation for Enviroments) [2], PMC (Parallel Model Combination) [3], RATZ (multivariate Gaussian based cepstral normalization) [4] and RASTA (the RelAtive SpecTrAl Technique) [5]. The research that followed this evolution was to make a proper combination of algorithms in order to reduce the noise effects. For example, a good example is described in [6], where the core scheme is composed of a Continuous SS (CSS) and PMC.

Persuing the same idea, a combination of the speech enhanced signal (represented by the SS method) and a feature vector normalization technique (PD-MEMLIN [7]) are presented in this work to improve the recognition accuracy of the speech recognition system in highly degraded environments [8, 9]. The first technique was selected because of its implementation simplicity and good performance. The second one is an empirical vector normalization technique that has been compared against some other algorithms [8] and has obtained important improvements.

The organization of the paper is as follows. In Section 2, a brief overview of the SS and PD-MEMLIN. Section 3 details the new method PD-MEEMLIN. In Section 4, the experimental results are presented. Finally, the conclusions are shown in Section 5.

2 Spectral Subtraction and PD-MEMLIN

In order to evaluate the proposed integration, an ASR system is employed. In general, a pre-processing stage of the speech waveform is always desirable. The speech signal is divided into overlaped short windows, from which a set of coefficients, usually Mel Frequency Cepstral Coefficients (MFCCs)[10], are computed. The MFCCs are feeded to the training algorithm that calculates the acoustic models. The acoustic models used in this research are the Hidden Markov Models (HMMs), which are widely used to model statistically the behaviour of the phonetic events in speech [10]. The HMMs employ a sequence of hidden states which characterises how a random process (speech in this case) evolves in time. Although the states are not observable, a sequence of realizations from these states can always be obtained. Associated to each state there is a probability density function, normally a mixture of Gaussians. The criteria used to train the HMMs is the Maximum Likelihood, thus, the training process becomes an optimization problem that can be solved iteratively with the Baum and Welch algorithm.

2.1 Spectral Subtraction

The Spectral Subtraction (SS) algorithm is a simple and known speech enhancement technique. This research is based on the SS algorithm expressed in [9]. It has the property that it does not requiere the use of an explicit voice activity detector, as general SS algorithms does. The algorithm is based on the existance of peaks and valleys in a short noisy speech time subband power estimate [9]. The peaks correspond to the speech activity and the valleys are used to obtain an estimate of the subband noise power. So, a reliable noise estimation is obtained using a large enough window that can pemit the detection of any peak of speech activity.

As shown in Figure 1, this algorithm performs a modification of the short time spectral magnitude of the noisy speech signal during the process of enhancement. Hence, the output signal can be considered close to the speech clean signal when

synthesized. The appropriate computation of the spectral magnitude is obtained with the noise power estimate and the SS algorithm. Let, $y(i) = x(i)+n(i)$, where

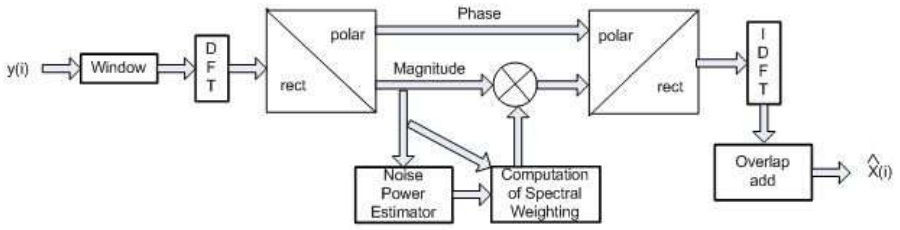

Fig. 1. Diagram of the Basic SS Method Used

$y(i)$ is the noisy speech signal, $x(i)$ is the clean speech signal, $n(i)$ is the noise signal and i denotes the time index, $x(i)$ and $n(i)$ are statistically independent.

Figure 1 depicts the spectral analysis in which the frames in the time domain data are windowed and converted to frequency domain using the Discrete Fourier Transform (DFT) filter bank with W_{DFT} subbands and with a decimation/interpolation ratio named R [9]. After the computation of the noise power estimation and the spectral weightening, the enhanced signal can be transformed back to the time domain using the Inverse Discrete Fourier Transform (IDFT).

For the subtraction algorithm it is necessary to estimate the subband noise power $P_n(\lambda, k)$ and the short time signal power $\overline{|Y(\lambda, k)|^2}$, where λ is the decimated time index and k are the frequency bins of the DFT. A first order recursive network is used to obtain a short time signal power as shown in Equation 1.

$$\overline{|Y(\lambda, k)|^2} = \gamma * \overline{|Y(\lambda - 1, k)|^2} + (1 - \gamma) * |Y(\lambda, k)|^2. \tag{1}$$

Afterwards, the subtraction algorithm is accomplished using an oversubtraction factor $osub(\lambda, k)$ and a spectral flooring constant ($subf$) [12]. The $osub(\lambda, k)$ factor is needed to eliminate the musical noise, and it is calculated as a function of the subband Signal to Noise Ratio $SNR_y(\lambda, k)$, λ and k (for a high SNR and high frequencies less $osub$ factor is required, for low SNR and low frequencies the $osub$ is less). The $subf$ constant helps the resultant spectral components from going below a minimum level. It is expressed as a fraction of the original noise power spectrum. The final relation of the spectral subtraction between $subf$ and $osub$ is defined by Equation 2.

$$|\hat{X}(\lambda, k)| = \begin{cases} \sqrt{subf * P_n(\lambda, k)} & \text{if } |Y(\lambda, k)| * Q(\lambda, k) \leq \sqrt{subf * P_n(\lambda, k)} \\ |Y(\lambda, k)| * Q(\lambda, k) & otherwise \end{cases} \tag{2}$$

where $Q(\lambda, k) = (1 - \sqrt{osub(\lambda, k) \frac{P_n(\lambda, k)}{|Y(\lambda, k)|^2}})$.
The missing element, $P_n(\lambda, k)$, is computed using the short subband signal power

$P_y(\lambda, k)$ in a representation based on smoothed periodograms, as denoted by $P_y(\lambda, k) = \xi * P_y(\lambda - 1, k) + (1 - \xi) * |Y(\lambda, k)|^2$ where ξ represents the smoothing constant to obtain the periodograms. Then, $P_n(\lambda, k)$ is calculated as a weighted minimum of $P_x(\lambda, k)$ in a window of D subband samples. Hence,

$$P_n(\lambda, k) = omin \cdot P_{min}(\lambda, k), \tag{3}$$

where $P_{min}(\lambda, k)$ denotes the estimated minimum power and $omin$ is a bias compensation factor. The data window D is divided into W windows of length M, allowing to update the minimum every M samples without time consuming. This noise estimator combined with the spectral subtraction has the ability to preserve weak speech sounds. If a short time subband power is observed, the valleys correspond to the noisy speech signal and are used to estimate the subband noise power.

The last element to be calculated is the $SNR_y(\lambda, k)$ in Equation 4 that controls the oversubtraction factor $osub(\lambda, k)$.

$$SNR_y(\lambda, k) = 10log\left(\frac{P_y(\lambda, k) - min(P_n(\lambda, k), P_y(\lambda, k))}{P_n(\lambda, k)}\right) \tag{4}$$

Up to this stage $osub(\lambda, k)$ and $subf$ can be selected and the spectral substraction algorithm can be computed.

2.2 PD-MEMLIN

PD-MEMLIN is an empirical feature vector normalization technique which uses stereo data in order to estimate the different compensation linear transformations in a previous training process. The clean feature space is modelled as a mixture of Gaussians for each phoneme. The noisy space is split in several basic acoustic environments and each environment is modelled as a mixture of Gaussians for each phoneme. The transformations are estimated for all basic environments between a clean phoneme Gaussian and a noisy Gaussian of the same phoneme.

PD-MEMLIN approximations Clean feature vectors, x, are modelled using a GMM for each phoneme, ph

$$p_{ph}(x) = \sum_{s_x^{ph}} p(x|s_x^{ph})p(s_x^{ph}), \tag{5}$$

$$p(x|s_x^{ph}) = N(x; \mu_{s_x^{ph}}, \Sigma_{s_x^{ph}}), \tag{6}$$

where $\mu_{s_x^{ph}}$, $\Sigma_{s_x^{ph}}$, and $p(s_x^{ph})$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with the clean model Gaussian s_x^{ph} of the ph phoneme.

Noisy space is split into several basic environments, e, and the noisy feature vectors, y, are modeled as a GMM for each basic environment and phoneme

$$p_{e,ph}(y) = \sum_{s_y^{e,ph}} p(y|s_y^{e,ph})p(s_y^{e,ph}), \tag{7}$$

$$p(y|s_y^{e,ph}) = N(y; \mu_{s_y^{e,ph}}, \Sigma_{s_y^{e,ph}}), \tag{8}$$

where $s_y^{e,ph}$ denotes the corresponding Gaussian of the noisy model for the e basic environment and the ph phoneme; $\mu_{s_y^{e,ph}}$, $\mathbf{\Sigma}_{s_y^{e,ph}}$, and $p(s_y^{e,ph})$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with $s_y^{e,ph}$.

Finally, clean feature vectors can be approximated as a linear function, f, of the noisy feature vector for each time frame t which depends on the basic environments, the phonemes and the clean and noisy model Gaussians: $x \approx f(y_t, s_x^{ph}, s_y^{e,ph}) = y_t - r_{s_x^{ph}, s_y^{e,ph}}$, where $r_{s_x^{ph}, s_y^{e,ph}}$ is the bias vector transformation between noisy and clean feature vectors for each pair of Gaussians, s_x^{ph} and $s_y^{e,ph}$.

PD-MEMLIN enhancement With those approximations, PD-MEMLIN transforms the Minimum Mean Square Error (MMSE) estimation expression, $\hat{x}_t = E[x|y_t]$, into

$$\hat{x}_t = y_t - \sum_e \sum_{ph} \sum_{s_y^{e,ph}} \sum_{s_x^{ph}} r_{s_x^{ph}, s_y^{e,ph}} p(e|y_t)p(ph|y_t, e)p(s_y^e|y_t, e, ph)p(s_x^{ph}|y_t, e, ph, s_y^e), \tag{9}$$

where $p(e|y_t)$ is the a posteriori probability of the basic environment; $p(ph|y_t, e)$ is the a posteriori probability of the phoneme, given the noisy feature vector and the environment; $p(s_y^{e,ph}|y_t, e, ph)$ is the a posteriori probability of the noisy model Gaussian, $s_y^{e,ph}$, given the feature vector, y_t, the basic environment, e, and the phoneme, ph. To estimate those terms: $p(e|y_t)$, $p(ph|y_t, e)$ and $p(s_y^{e,ph}|y_t, e, ph)$, (7) and (8) are applied as described in [8]. Finally, the cross-probability model, $p(s_x^{ph}|y_t, e, ph, s_y^{e,ph})$, which is the probability of the clean model Gaussian, s_x^{ph}, given the feature vector, y_t, the basic environment, e, the phoneme, ph, and the noisy model Gaussian, $s_y^{e,ph}$, and the bias vector transformation, $r_{s_x^{ph}, s_y^{e,ph}}$, are estimated in a training phase using stereo data for each basic environment and phoneme [8].

3 PD-MEEMLIN

By combinig both techniques, PD-MEEMLIN arises as an empirical feature vector normalization which estimates different linear transformations as PD-MEMLIN, with the special property that a new enhanced space is obtained by applying SS to the noisy speech signal. Furthermore, this first-stage enhancement produces that the noisy space gets closer to the clean one, making the gap smaller among them. Figure 2 shows PD-MEEMLIN architecture.

Next, the architecture modules are explained:

– The SS-enhancement of the noisy speech signal is performed, $|\hat{X}(\lambda, k)|$, $P_n(\lambda, k)$ and $SNR_y(\lambda, k)$ are calculated.
– Given the clean speech signal and the enhanced noisy speech signal, the clean and noisy-enhanced GMMs are obtained.
– In the testing stage, the noisy speech signal is also SS-enhanced and then normalized using PD-MEEMLIN.
– These normalized coefficients are forwarded to the decoder.

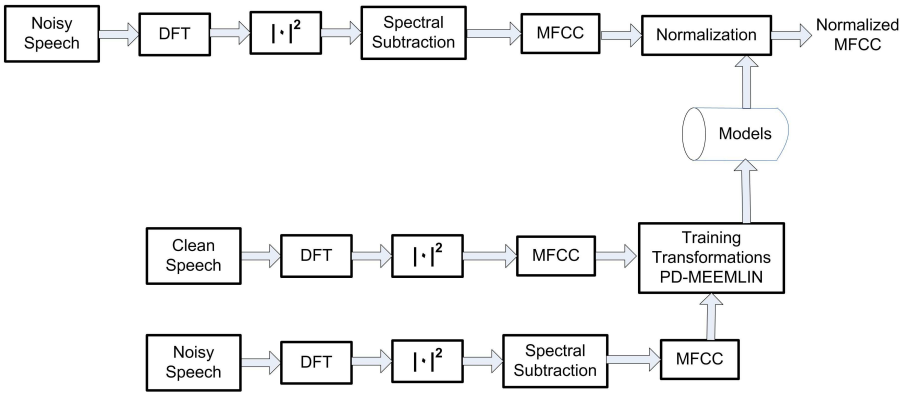

Fig. 2. PD-MEEMLIN Architecture.

4 Experimental Results

All the experiments were performed employing the AURORA2 database [13], clean and noisy data based on TIDigits. Three types of noises were selected: Subway, Babble and Car from AURORA2, that go from -5dB to 20dB SNR. For every SNR the SS parameters *osub* and *subf* needs to be configured. The parameter *osub* takes values from 0.4 to 4.6 (0.4 for 20dB, 0.7 for 15dB, 1.3 for 10dB, 2.21 for 5dB, 4.6 for 0dB and 4.6 for -5dB) and *subf* values 0.03 or 0.04 (all SNR levels except 5dB optimised for 0.04). The phonetic acoustic models employed by PD-MEEMLIN are obtained from 22 phonemes and 1 silence. The models set is represented by a mixture of 32 Gaussians each. Besides, two new sets of each noise were used, PD-MEEMLIN needs one to estimate the enhanced-noisy model, and onother to obtain the normalized coefficients. The feature vectors for the recognition process are built by 12 normalized MFCCs followed by the energy coefficient, its time-derative Δ and the time-acceleration $\Delta\Delta$. For the training stage of the ASR system, the acoustic models of 22 phonemes and the silence consist on a three-state HMMs with a mixture of 8 Gaussians per state. The combined techniques show that for low noise conditions i.e. SNR=10, 15 or

20 dB, the difference between the original noisy space and the one approximated to the clean is similar. However, when the SNR is lower (-5dB or 0dB) the SS improves the performance of PD-MEMLIN. Comparing the combination of SS with PD-MEMLIN against the case where no techniques are applied, a significant improvement is shown. The results described before are presented in Tables 1, 2 and 3. The Tables show "Sent" that means complete utterances percent-

Table 1. Comparative Table for the ASR working with Subway Noise.

Subway	ASR		ASR+SS		ASR+PD-MEMLIN		ASR+PD-MEEMLIN	
SNR	**Sent %**	**Word %**	**Sent %**	**Word %**	**Sent %**	**Word %**	**Sent %**	**Word %**
-5dB	3.40	21.57	10.09	34.22	11.29	37.09	13.29	47.95
0dB	9.09	29.05	20.18	53.71	27.07	61.88	30.87	69.71
5dB	17.58	40.45	32.17	70.00	48.15	80.38	51.65	83.40
10dB	33.07	65.47	50.95	83.23	65.83	90.58	70.13	91.86
15dB	54.45	84.60	64.84	90.02	78.92	94.98	78.22	94.40
20dB	72.83	93.40	76.52	94.56	85.91	97.14	86.71	97.30

Table 2. Comparative Table for the ASR working with Babble Noise.

Babble	ASR		ASR+SS		ASR+PD-MEMLIN		ASR+PD-MEEMLIN	
SNR	**Sent %**	**Word %**	**Sent %**	**Word %**	**Sent %**	**Word %**	**Sent %**	**Word %**
-5dB	4.60	23.08	7.59	29.78	8.49	29.54	6.69	37.79
0dB	11.29	30.41	15.98	44.49	23.48	55.72	20.08	59.50
5dB	20.58	44.23	30.37	65.11	48.75	80.55	49.25	83.70
10dB	40.86	72.85	50.25	80.93	74.93	94.20	69.33	91.48
15dB	69.03	90.54	69.93	90.56	84.12	96.86	81.32	95.54
20dB	82.42	96.17	83.52	95.84	88.91	98.09	88.01	97.98

Table 3. Comparative Table for the ASR working with Car Noise.

Car	ASR		ASR+SS		ASR+PD-MEMLIN		ASR+PD-MEEMLIN	
SNR	**Sent %**	**Word %**	**Sent %**	**Word %**	**Sent %**	**Word %**	**Sent %**	**Word %**
-5dB	3.10	20.18	10.49	28.87	6.79	25.90	13.89	44.31
0dB	8.09	26.18	18.58	46.70	23.58	52.67	35.16	70.47
5dB	14.99	35.34	31.47	66.50	51.95	82.34	58.64	86.30
10dB	28.77	58.13	54.25	82.72	70.83	92.15	70.93	91.90
15dB	57.84	84.04	68.03	90.51	82.02	96.16	81.42	95.86
20dB	78.32	94.61	81.42	95.30	87.01	97.44	87.81	97.77

age correctly recognised, and "Word" indicates the words percentage correctly

recognised. The gap between the clean and the noisy model, for the very high degraded speech, had been shortened due to the advantages of both techniques. When PD-MEEMLIN is employed the performance is between 11.7% and 24.84% better than PD-MEMLIN, and between 11.4% and 34.5% better than SS.

5 Conclusions

In this work a robust normalization technique, PD-MEEMLIN, has been presented by cascading a speech enhancement method (SS) followed by a feature vector normalization algorithm (PD-MEMLIN). The results of PD-MEEMLIN show a better performance than SS and PD-MEMLIN for a very high degraded speech. This improvement is made by the enhancement of the noisy models employed by PD-MEMLIN, which are close to the original clean model.

References

1. S. Boll: Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Trans ASSP, Vol27, pp. 113–120, 1979.
2. J. Droppo, L. Deng, and A. Acero: Evaluation of the Splice Algorithm on the Aurora2 Database. In Proc. Eurospeech, Vol. 1, Sep. 2001.
3. M.J.F. Gales, and S. Young: Cepstral Parameter Compensation for HMM Recognition in Noise. Speech Communication, Vol. 12 Issue 3, pp. 231–239, 1993.
4. Pedro J. Moreno, Bhiksha Raj, Evandro Gouvea and Richard M. Stern: Multivariate-Gaussian-Based Cepstral Normalization for Robust Speech Recognition. Department of Electrical and Computer Engineering & School of Computer Science. Carnegie Mellon University.
5. Hynek Hermansky and Nelson Morgan: RASTA Processing of Speech. IEEE Transactions on Speech and Audio Processing, Vol. 2 No. 4, pp. 578–589, October 1994.
6. J. Nolazco-Flores and S. Young: Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM adaptation. In ICASSP, pages I.409–I.412 (1994).
7. L. Buera, E. Lleida, A. Miguel, and A. Ortega: Multienvironment Models Based LInear Normalization for Speech Recognition in Car Conditions. Proc. ICASSP, May 2004.
8. L. Buera, E. Lleida, A. Miguel, and A. Ortega: Robust Speech Recognition in Cars Using Phoneme Dependent Multienvironment LInear Normalization. In Proceedings of Interspeech. Lisboa, Portugal, 2005, pp. 381-384.
9. R. Martin: Spectral Subtraction Based on Minimum Statistics. In Proc. Eur. Signal Processing Conf. 1994, pp. 1182-1185.
10. Xuedong Huang, Alex Acero and Hsiao-Wuen Hon: Spoken Language Processing. Prentice Hall PTR, United States, pp. 504–512, 2001.
11. R. Martin: Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics. IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 5, July 2000.
12. M. Berouti, R. Schwartz, and J. Makhoul: Enhancement of Speech Corrupted by Acoustic Noise. Proc. IEEE Conf. ASSP, pp. 208-211, April 1979.
13. H. G. Hirsch and D. Pearce: The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems Under Noisy Condidions. In ISCA ITRW ASR2000, Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, September 2000.