FEATURE EXTRACTION COMBINING SPECTRAL NOISE REDUCTION AND CEPSTRAL HISTOGRAM EQUALIZATION FOR ROBUST ASR

J.C. Segura, M.C. Benítez, A. de la Torre, A.J. Rubio

Dpto. de Electrónica y Tecn. de Comp. Universidad de Granada, 18071-Granada, SPAIN
{segura,carmen,atv,rubio}@ugr.es

ABSTRACT

This work is mainly focused on showing experimental results using a combination of two methods for noise compensation which are shown to be complementary: classical spectral subtraction algorithm and histogram equalization. While spectral subtraction is focused on the reduction of the additive noise in the spectral domain, histogram equalization is applied in the cepstral domain to compensate the remaining non-linear effects associated to channel distortion and additive noise. The estimation of the noise spectrum for the spectral subtraction method relies on a new algorithm for speech / non-speech detection (SND) based on order statistics. This SND classification is also used for dropping long speech pauses. Results on Aurora 2 and Aurora 3 are reported.

1. INTRODUCTION

In previous works [1, 2, 3] we have shown that histogram equalization (HE) is a very suitable tool for compensating the linear and non-linear distortions introduced into the speech signal by the environment (noise and channel distortions) in the cepstral domain. We have found that this technique can also be used in combination with noise reduction techniques. For example, the combination of Vector Tailor Series (VTS) (working in the filter-bank domain) [4] and HE (performed in the cepstral domain) improves the recognition performance with respect to each technique independently applied [1].

In this work we show that it is possible to use histogram equalization in combination with another classical noise reduction method like spectral subtraction (SS). The main advantage of using SS instead of VTS is that SS does not need a model of the clean speech signal. Additionally the computational complexity is reduced. The goal of applying HE after SS (or other noise reduction techniques like VTS or Wiener filter) is to remove the mismatch between the clean speech and the partially compensated speech; that means that histogram equalization is applied to a signal with a signal to noise ratio (SNR) higher than that of the original one because the noise was partially removed.

Figure 1 shows a block diagram of the proposed feature extraction algorithm. In the context of Distributed Speech Recognition systems (DSR) the proposed techniques can be distributed between the front-end and the back-end. At the front-end, spectral subtraction is used to provide a noise-reduced speech signal. A Speech / Non-speech Detector (SND) module using logarithmic energy classifies frames as speech or non-speech for the estimation of the noise spectrum (NS). This estimation is based on the magnitude spectrum of those frames classified as non-speech by the SND algorithm, and it is used to perform classical spectral subtraction. Finally 12 MFCC coefficients and the logarithmic energy are computed from the noise-reduced signal.

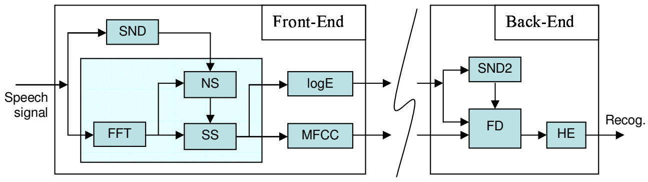

Fig. 1. *Block diagram of the proposed system.*

At the back-end, a second SND (SND2) is used by a frame-dropping algorithm (FD) to remove long speech pauses. The goal of this module is to reduce the insertion error rate. After that, histogram equalization over the cepstrum coefficients is performed to remove the remaining distortion in the speech representation.

For the design of the SND module we propose a new approach based on order statistics (OS) [5, 6] to obtain an estimation of the instantaneous SNR of the speech signal, that is used for the speech / non-speech detection.

This paper is organized as follows. Section 2 describes the noise reduction algorithm including the new approach for speech / non-speech detection. In section 3 we describe the implementation used for histogram equalization. In section 4 the speech data bases are described and the experimental results are discussed. Finally, in section 5 the main conclusions are presented.

2. FRONT-END NOISE REDUCTION

2.1. Noise estimation and spectral subtraction

Noise reduction in the front-end is based on an implementation of the traditional non linear spectral subtraction algorithm in the magnitude spectrum domain [7]:

$$|\hat{X}| = \max\{(|Y| - \alpha\overline{|N|}), \beta|Y|\} \tag{1}$$

where $|\hat{X}|$ is the compensated magnitude spectrum, $|Y|$ is the magnitude spectrum of the noisy signal, $\overline{|N|}$ is the average magnitude spectrum of the noise, $\alpha = 1.1$ is the over-subtraction factor and $\beta = 0.3$ is the maximum attenuation.

To obtain an estimation of the noise spectrum, we use the information provided by the SND module. Only when the SND classifies the current frame as non-speech the noise spectrum is adapted using a first order IIR filter with a forgetting factor $\lambda = 0.95$ as follows:

This work has been supported by the Spanish Government under the CICYT project TIC2001-3323.

$$\overline{|N_t|} = \begin{cases} \lambda\overline{|N_{t-1}|} + (1-\lambda)|Y_t|; & \text{if frame } t \text{ is silence} \\ \overline{|N_{t-1}|}; & \text{if frame } t \text{ is speech} \end{cases} \tag{2}$$

2.2. Speech / Non-speech Detection

The detection of speech pauses is a difficult task particularly when the SNR is low. In this work we propose a new approach to speech pause detection based on order statistics (OS) filters [6]. These are a special class of nonlinear filters that can be viewed as a generalization of the median filter. We apply this theory to obtain an estimation of the local SNR of the speech signal. Two OS filters are applied to the log energy of the signal. The first one is a median filter used to track the background noise level (B). The other one is used to track the signal level and it is defined as the 0.9 quantile ($Q(0.9)$). The difference between the output of this filter and the background noise level is used as a quantile-based estimation of the instantaneous SNR (QSNR) of the signal.

The implementation of these two OS filters is based on the sequence of sorted values of log energy. Let $E_{t-L} \cdots E_{t+L}$ be the log energy values of $2L+1$ frames around the frame t to be analyzed. Let $E_{(r)}$, where $r = 1 \cdots 2L+1$, be the corresponding sorted values in ascending order. Then, $E_{(L+1)}$ is the output of the median filter. For the other filter we use the general expression

$$\begin{aligned} r = 2pL \quad k = \lfloor r \rfloor \quad f = (r-k) \\ Q(p) = (1-f)E_{(k)} + fE_{(k+1)} \end{aligned} \tag{3}$$

where $Q(p)$ is the p sampling quantile and $\lfloor r \rfloor$ denotes the greatest integer smaller than r. The algorithm can be summarized as follows:

1. Log energy is computed for 20ms frames at a frame-rate of 100Hz.
2. Two OS filters of length $2L+1$ are used to obtain estimations of the median and the quantile $Q(0.9)$ of the logarithmic energy.
3. An approximation to the local SNR (QSNR) of the signal in the working window is obtained as the difference between $Q(0.9)$ and the background noise level B.
4. The speech / non-speech detection is made by comparing the estimated SNR with a threshold. If QSNR is greater than the threshold the frame is classified as speech, and otherwise it is classified as non-speech; in this last case the background noise level is updated using the median value obtained for this window.

For the initialization of the algorithm we consider the first L frames as non-speech, and the median of this L frames is used as an initial estimation of the background noise level. In this work we use a 21 frames window ($L = 10$) and a threshold value of 3dB.

Figure 2 shows how this algorithm works. The first picture represents a speech signal and the output of the SND. The second one shows the signals involved in the decision algorithm; at the top the logarithmic energy, the quantile $Q(0.9)$ and the background noise level B; at the bottom the SNR estimate (QSNR) and the decision threshold.

More traditional algorithms for speech / non-speech detection based on energy values obtain an estimation of the instantaneous SNR as the difference between the log energy for the current frame and an estimation of the log energy of the noise. Usually some type of smoothing and hand-over is applied. In our approach, this is not necessary. On the picture we can observe that QSNR is smooth enough and performs an advanced detection of noise/speech transitions and a delayed detection of speech/noise transitions; this occurs because of the difference delay behavior of non-symmetric OS filters for non-increasing and non-decreasing sequences [6].

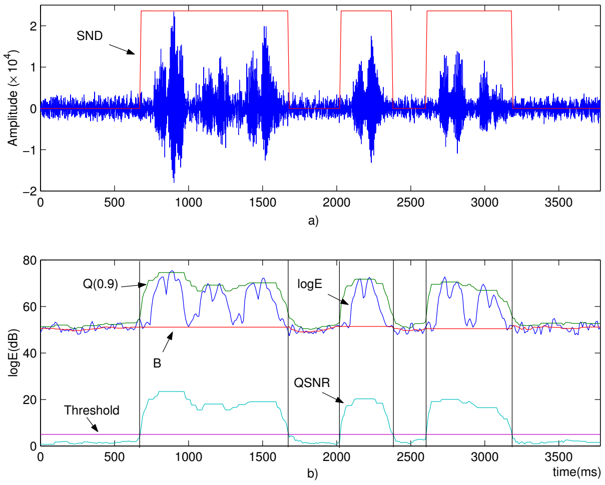

Fig. 2. *(a) Speech signal and speech / non-speech detection. (b) At the top the logE, the quantile $Q(0.9)$ and the noise level estimation (B); at the bottom the SNR estimate (QSNR) and the threshold value.*

3. BACK-END FEATURE COMPENSATION

3.1. Frame-dropping

At the back-end, a second SND (SND2) is applied over the log energy of the noise-reduced signal (after the SS is applied). The output of this module is used in a frame-dropping algorithm with the goal of removing long speech pauses. It removes every frame in the middle of a window of 11 non-speech frames. Consequently, the maximum length of speech pauses allowed after the frame-dropping is 10 frames (100ms).

3.2. Histogram Equalization

Histogram equalization is used to compensate both, linear and non-linear distortions of the feature vector. Using spectral subtraction, a reduction of the mismatch between noisy and clean speech is obtained. Nevertheless, this compensation is not perfect and a residual mismatch still remains. In addition, channel mismatch distortion is not removed by spectral subtraction at all. The goal of HE is to further reduce this two residual mismatches. In [1] the nature of this residual noise was studied when VTS noise compensation is applied. It was found that histogram equalization is a good tool to deal with it.

Histogram equalization is a non linear technique originally developed for image processing [8] that has been applied with good results to noise compensation in ASR systems [3]. The goal of this technique is to provide a function $x(y)$ which transforms the probability distribution of the noisy speech $p_y(y)$ into a reference probability distribution corresponding to clean speech $p_x(x)$. If $x(y)$ transforms $p_y(y)$ into $p_x(x)$, then the cumulative histograms verify that [8]

$$C_y(y) = C_x(x(y)) \tag{4}$$

and therefore, the transformation $x(y)$ providing an estimation of the clean speech can be obtained from the cumulative histograms of the noisy and clean speech

$$x(y) = C_x^{-1}[C_y(y)] \tag{5}$$

where C_x^{-1} represents the inverse function of C_x.

Although it is possible to use reference histograms estimated from clean speech samples, the considered reference probability density function in this work is a Gaussian with zero mean and unity variance. We apply histogram equalization in the cepstrum domain exactly in the same way that is described in [1].

4. EXPERIMENTAL RESULTS

4.1. Speech Databases and setup

Experiments were performed on 4 different databases: Aurora 2 [9] and Finnish [10], Spanish [11] and German [12] SpeechDat Car databases.

Aurora 2 database is based on a subset of the TI-Digits database. This database contains connected digits recorded in a clean environment. Several types of noises (subway, babble, car, exhibition, etc.) at different SNR levels are added to the sentences. Three sets of sentences for test (A, B and C), and two groups of recognition experiments, one using a recognizer trained with clean speech (Clean-Condition) and other trained with sentences contaminated with different kinds and levels of noise (Multi-condition) are defined for this database [9].

The SpeechDat Car databases were recorded in a car environment in several driving conditions with two microphones (close talking and hands free). Three sets of experiments with increasing level of mismatch between training and test conditions were defined: well matched (WM), medium mismatch (MM) and high mismatch (HM). The three databases contain only digits utterances.

These databases have been automatically end-pointed. We have used the SND previously described to detect the beginning and the end of each sentence using close talking microphone recordings. Then we have added 200ms of silence at the beginning and at the end of each sentence, and we have used these endpoints also for hands free microphone recordings.

The reference recognition system [9] is based on continuous Hidden Markov models (one model for each digit) with 16 emitting states and a mixture of 3 Gaussian pdf per state. Both, training and recognition processes are performed using the HMM Tool Kit (HTK) [13] software, as proposed in Aurora 2 and Aurora 3 documentation.

4.2. Results and discussion

In this work we present the speech recognition results obtained with the baseline system based on MFCC and three different sets of features: SS features (obtained using only the spectral subtraction noise reduction algorithm at the front-end), SS+HE features (obtained after histogram equalization at the back-end) and SS+FD+HE features (similar to SS+HE but including frame - dropping before HE). In all cases, the feature vector (containing 12 MFCC and log energy) is augmented with its corresponding derivatives and accelerations using regression lengths of 7 and 11 frames respectively.

TI-Digits Multi-condition Training				
	A	B	C	Average
Baseline	88.07	87.22	84.56	87.03
SS	90.94	88.69	86.29	89.11
SS+HE	90.72	89.74	90.03	90.19
SS+FD+HE	90.89	89.80	90.11	90.30
SS+FD+HE (20mix)	91.97	91.62	90.46	91.53
TI-Digits Clean-Condition Training				
	A	B	C	Average
Baseline	58.74	53,40	66.00	58.06
SS	73.71	69.35	75.63	72.35
SS+HE	82.08	82.61	81.73	82.22
SS+FD+HE	82.51	82.78	81.87	82.49
SS+FD+HE (20mix)	82.89	84.03	82.27	83.22

Table 1. *Word accuracy results for TI-Digits*

Finnish				
	WM	MM	HM	Average
Baseline	92.74	80.51	40.53	75.41
SS	95.09	78.80	69.19	82.91
SS+HE	94.58	86.53	74.20	86.67
SS+FD+HE	94.58	86.73	73.11	86.46
Spanish				
	WM	MM	HM	Average
Baseline	92.94	83.31	51.55	79.22
SS	95.58	89.76	71.94	87.63
SS+HE	96.15	93.15	86.77	93.00
SS+FD+HE	96.65	94.10	87.03	93.35
German				
	WM	MM	HM	Average
Baseline	91.20	81.04	73.17	83.14
SS	93.41	86.60	84.32	88.75
SS+HE	94.79	88.58	89.32	91.25
SS+FD+HE	94.57	88.07	88.95	90.89

Table 2. *Word accuracy results for SpeechDat Car databases*

Tables 1 and 2 show the word accuracies for the different databases and test conditions for the three types of features. For TI-Digits, results for each test set are averaged over the SNR levels between 20dB and 0dB. Average values for SpeechDat Car databases are weighted as 0.4·WM + 0.35·MM + 0.25·HM.

Tables 3, 4 and 5 show the relative performance improvements over the proposed baseline. For the three systems, a relative improvement, in average, is obtained for all the databases, with the exception in the Finnish database medium mismatched set when only spectral subtraction is applied; in this case a degradation respect to the baseline is observed.

Note the important increment of the relative improvement achieved after combining spectral subtraction and histogram equalization (from 23.57% and 30.54% to 35.51% and 45.79% for TI-Digits and SpeechDat Car, respectively). This improvement is obtained for all the different databases and test conditions.

Relative improvement is increased for TI-Digits and Spanish

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	13,14%	7,42%	6,04%	9,43%
Clean	38,58%	39,65%	32,08%	37,71%
Average	25,86%	23,53%	19,06%	23,57%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	32,37%	37,39%	25,11%	▨	31,63%
Mid (x35%)	-8,77%	38,65%	29,32%	▨	19,73%
High (x25%)	48,19%	42,08%	41,56%	▨	43,94%
Overall	21,92%	39,00%	30,70%	▨	30,54%

Table 3. *Results obtained applying SS*

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	9,85%	17,51%	22,36%	15,42%
Clean	50,15%	64,27%	49,13%	55,59%
Average	30,00%	40,89%	35,74%	35,51%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	25,34%	45,47%	40,80%	▨	37,20%
Mid (x35%)	30,89%	58,96%	39,77%	▨	43,20%
High (x25%)	56,62%	72,69%	60,19%	▨	63,17%
Overall	35,10%	57,00%	45,29%	▨	45,79%

Table 4. *Results obtained combining SS and HE*

SpeechDat Car after applying frame-dropping. This is mainly due to the reduction of insertion errors caused by the long inter-word speech pauses removed by the FD algorithm. Nevertheless including frame-dropping does not improve performance in Finnish and German databases. It is necessary to recall that SpeechDat Car databases have been well delimited and for this reason the effect of frame dropping is not very relevant. Experiments performed without an a-priori delimitation of the databases have shown the importance of the frame-dropping because it reduces considerably the number of insertion error. Additionally, we have also performed recognition experiments for TI-Digits increasing the number of Gaussians from 3 to 20. Table 6 shows the relative improvements of the SS-FD-HE recognizer using 20 Gaussians with respect to the 3 Gaussians baseline system. As expected, the performance improvement is greater for Multi-Condition than for Clean-Condition, yielding to an average relative improvement of 52.02% over the baseline.

5. CONCLUSION

The recognition results on Aurora 2 and Aurora 3 show the effectiveness of the combination of spectral subtraction and histogram equalization. After applying spectral subtraction in the magnitude spectrum domain an important reduction of the mismatch is obtained, yielding to a performance increment over the baseline. Recognition accuracy can be further improved by applying histogram equalization in the cepstral domain. This is due to the ability of histogram equalization to compensate the linear channel distortion and the residual non-linear distortions remaining after spectral noise reduction. Frame-dropping is effective only in situations where long speech pauses occur in the utterances or when the beginning or the end of the speech utterance is not accurately delimited.

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	14,07%	18,31%	25,16%	17,99%
Clean	51,98%	64,20%	49,89%	56,45%
Average	33,03%	41,26%	37,53%	37,22%

Aurora 3 Relative Improvement					
	Finnish	Spanish	German	Danish	Average
Well (x40%)	25,34%	52,55%	38,30%	▨	38,73%
Mid (x35%)	31,91%	64,65%	37,08%	▨	44,55%
High (x25%)	54,78%	73,23%	58,81%	▨	62,28%
Overall	35,00%	61,95%	43,00%	▨	46,65%

Table 5. *Results obtained combining SS, FD and HE*

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	38,95%	43,35%	42,30%	41,38%
Clean	58,02%	70,98%	55,36%	62,67%
Average	48,48%	57,16%	48,83%	52,02%

Table 6. *Results obtained combining SS, FD and HE for the 20 Gaussian mixture system*

6. REFERENCES

[1] J.C. Segura, C. Benitez, A. de la Torre, S. Dupont, A.Rubio. *VTS Residual noise compensation. Accepted for ICASSP 2002*, Session: Robust Speech Recognition II, 2002.

[2] A. de la Torre. *Técnicas de mejora de la representación en los sistemas de Reconocimiento Automático de Voz*. PhD thesis, Universidad de Granada, España, April 1999.

[3] A. de la Torre, J.C. Segura, C. Benitez, A.M. Peinado and A.J. Rubio. *Non-linear transformations of the feature space for robust speech recognition. Accepted for ICASSP 2002*, Session: Robust Speech Recognition II, 2002.

[4] P. J. Moreno. *Speech Recognition in Noise Environments*. PhD thesis, University of Pittsburgh, Pensilvania, April 1996.

[5] A. Stuart, J. Keith, M. Kendall. *Kendall's advanced theory of statistics, 5^{th} Ed.*. Charles Griffin, London, 1987.

[6] H.G. Longbotham, A.C. Bovik. *Theory of Order Statistic Filters and their Relationship to Linear FIR Filters* IEEE Trans. on Acoustics, Speech, and Signal Processing. Vol. 37, NO. 2, February 1989.

[7] Saeed V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction*. Wiley, 2000.

[8] J.C. Russ. *The image processing handbook*. CRC Press, 1995.

[9] H.G. Hirsch and D. Pearce. *The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions. ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, Paris, France, September 2000.

[10] Nokia *Availability of Finnish SpeechDat Car database for ETSI STQ WI008 front-end standarization*. November, 1999.

[11] D. Macho *Spanish SDC-AuroraDatabase for ETSI STQ Aurora WI008 Advanced DSR front-end and evaluation: Description and Basline Results*. December, 2000.

[12] L. Netsch. *Description and Baseline Results for the subset of the Speechdat-Car German database used for ETSI STQ AURORA WI008 Advanced DSR Evaluation*. January, 2001.

[13] S. Young, J. Odell, D. Ollason, V. Valtchev and P. Woodland. *The HTK Book*. Cambridge University, 1997.