NOISE SUPPRESSION AND LOUDNESS NORMALIZATION IN AN AUDITORY MODEL-BASED ACOUSTIC FRONT-END

Halewijn Vereecken and Jean-Pierre Martens [1]

ELIS, University of Ghent
St.-Pietersnieuwstraat 41
B-9000 Gent (Belgium)
halewijn/martens@elis.rug.ac.be

ABSTRACT

It is commonly acknowledged that the presence of additive and convolutional noise and speech level variations can seriously deteriorate the performance of a speech recognizer. In case an auditory model is used as the acoustic front-end, it turns out that compensation techniques such as spectral subtraction and log-spectral mean subtraction can be outperformed by time-domain techniques operating on the band-pass filtered signals which are supplied to the haircell models. In [1] we showed that additive noise could be removed effectively by means of center clippers put in front of the haircell models. This technique, which was called linear noise magnitude subtraction (NMS), is further improved in this paper. The nonlinear NMS proposed here outperforms the linear one, especially for low Signal-to-Noise Ratios. To compensate for speech level variations and convolutional noise, we have adopted the same filosophy: remove the effects before the signal is supplied to the haircell models. This is accomplished by introducing normalization gains in front of the haircell models. It is shown that this loudness mean normalization (LMN) technique when used in combination whith NMS offers a highly robust speech representation.

1. INTRODUCTION

Speech recognition is concerned with retrieving the linguistic message, given the acoustic observations. In practice, the environment in which the observations are obtained, may differ severely from the one in which the training examples were gathered. In this paper we focus on three factors contributing significantly to the performance degradation of a speech recognizer (figure 1):

- *Additive noise* such as car noise and noise from interfering speakers.
- *Convolutional noise* emerging from room acoustics or recording equipment, e.g. telephone channel.
- *Speech level variations* arising from the speaker (weak or strong voice), changes of orientation or distance to the microphone, or unknown attenuation in the recording channel.

Other distortions include nonlinearities due to switching on the telephone network, and the effects of noise and stress on the speaking style. Although these effects may prove to be equally important, they will not be considered here.

In a previous paper, we showed that inserting a center clipper between the cochlear filter and the haircell in each channel of an auditory model offers more robustness to additive noise than a generalized power spectral subtraction applied to the outputs of these channels. As the clipping level was chosen proportional to the estimated noise magnitude in the channel, the technique was called linear noise magnitude subtraction. In this paper, nonlinear noise magnitude subtraction is introduced. It significantly outperforms the original algorithm for low SNRs.

In order to deal with convolutional noise and speech level variations, we have introduced a loudness normalization gain between the cochlear filter and the center clipper. The aim is to make the average loudness in each channel independent of the input level in that particular channel. Experiments show that the loudness mean normalization definitely improves the performance of the recognizer.

2. NOISE MAGNITUDE SUBTRACTION

Previously [1], we have shown that additive noise can be removed by performing a noise magnitude subtraction (NMS) in each channel of an auditory model. No prior information about the noise (e.g. spectrum or noise level) is required, and the recognizer itself need not be altered, although error rates can be reduced further by retraining the phone models. The NMS is performed by a center clipper (figure 2), whose clipping level $\Delta(n)$ at time n is set equal to twice the standard deviation of the noise. The latter is derived from the minimum value of $q(n)$ (the variable controlling the gain of the haircell model) measured in a window of 1.5 seconds preceding n.

2.1. Nonlinear NMS

The idea of nonlinear NMS is similar to that used in nonlinear spectral subtraction [2]: remove less noise in high and more noise in low Signal-to-Noise Ratio (SNR) regions. For spectral subtraction however, this lead to the use of rather ad hoc subtraction functions. By demanding that the clipped noise causes a fixed but small average response at the output of the channel, we could find an analytical solution. If we take into account that the noise $x(n)$ (figure 2) exhibits a zero-mean Gaussian density function with deviation σ, the expected

[1] Research Associate of the National Fund for Scientific Research

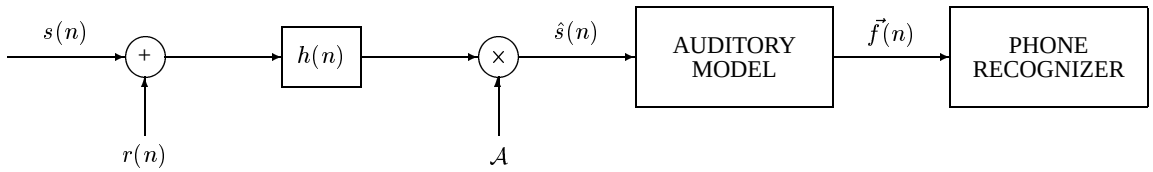

Figure 1: Noise free speech $s(n)$ may be distorded by additive noise $r(n)$ and by a linear filter with impulse response $h(n)$. Speech level variations (modelled by $\mathcal{A}$) may occur due to line attenuation, distance to microphone, ...

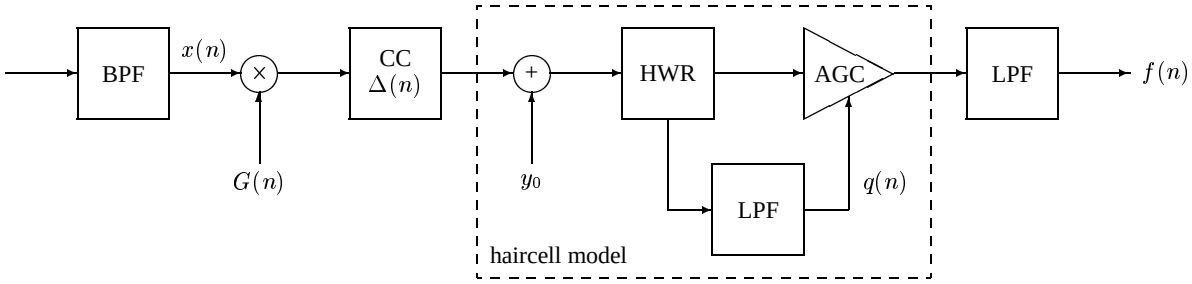

Figure 2: Each channel of the auditory model incorporates a band-pass filter (BPF), a variable gain $G(n)$, a center clipper (CC) with clipping level $\Delta(n)$, an offset y_0, a halfwave rectifier (HWR), an automatic gain control device (AGC) and a low-pass filter (LPF) to extract the firing rates $f(n)$.

value of $q(n)$ can be calculated as

$$\overline{q} = \frac{y_0 - \Delta}{2} + \frac{y_0 + \Delta}{2}\,\Phi_1(\gamma) = y_0 + \frac{G\sigma}{\sqrt{2}}\,\Phi_2(\gamma)$$

with

$$\begin{aligned}
\gamma &= \frac{y_0 + \Delta}{G\sigma\sqrt{2}} \\
\Phi_1(\gamma) &= \frac{1}{\sqrt{\pi}}\frac{e^{-\gamma^2}}{\gamma} + \mathrm{erf}(\gamma) \\
\Phi_2(\gamma) &= \gamma\,[\Phi_1(\gamma) - 1] \\
\mathrm{erf}(\gamma) &= \frac{2}{\sqrt{\pi}}\int_0^{\gamma} e^{-t^2}\,dt
\end{aligned}$$

If Δ_1 is the current value of the clipping level and $\overline{q}_1$ the current value of $\overline{q}$, γ_1 can be calculated as

$$\gamma_1 = \Phi_1^{-1}(\phi_1) \quad \text{with} \quad \phi_1 = \frac{2\,\overline{q}_1 - y_0 + \Delta_1}{y_0 + \Delta_1}$$

and the deviation of the noise $x(n)$ is obtained as

$$\sigma = \frac{y_0 + \Delta_1}{G\,\gamma_1\sqrt{2}}$$

By demanding that with the new Δ_2, $\overline{q}$ would become equal to $y_0 + \epsilon\,y_0$, γ_2 is obtained as

$$\gamma_2 = \Phi_2^{-1}(\phi_2) \quad \text{with} \quad \phi_2 = \frac{\epsilon\,y_0\sqrt{2}}{G\sigma}$$

and the next value of the clipping level becomes

$$\Delta_2 = \max\left(G\sigma\,\gamma_2\sqrt{2} - y_0\,,\,0\right)$$

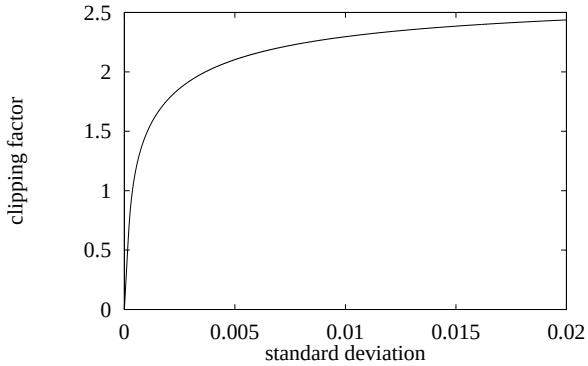

Figure 3: Clipping factor as a function of $G\sigma$ ($\epsilon = 0.03$).

Both $\Phi_1(\gamma)$ and $\Phi_2(\gamma)$ are monotonously decreasing, and can easily be tabulated, so that solving their inverse introduces no computational overhead. Figure 3 depicts $\gamma_2\sqrt{2}$ as a function of $G\sigma$. Clearly, the clipping level is a nonlinear function of the noise magnitude.

An estimate of $\overline{q}$ is obtained by measuring the minimum of $q(n)$ in a window of 1260 ms. This window is considered large enough to bridge at least one silence or closure in the signal. However, since the average time constant of the haircell model LPF is not larger than 12 ms, this minimum would be an underestimate of $\overline{q}$. Therefore we have averaged a number of subsequent samples of $q(n)$ in a time interval of 36 ms. Furthermore, it is important to observe that it will take some time before the effect of the abrupt change of $\Delta(n)$ from Δ_1 to Δ_2 will be fully effective in $q(n)$. Therefore, the clipping level is frozen to Δ_2 for the next 90 ms. In general, if a *new* minimum is found, and it is larger than y_0, the procedure described above

is followed. If the new minimum is smaller than y_0, ϕ_1 is smaller than 1 and no γ_1 exists. Then, $\Delta(n)$ is replaced by $0.75\Delta(n)$.

2.2. Initial Noise Estimate

Previously it took at least 150 ms initial noise to get a reliable initial Δ. In fact, if the noise fragment is shorter, the probability density function of the estimate becomes wider, especially if there is a lot of noise in the channel. However, in that case some of the speech will be masked, and this part of the signal can thus be used to refine the estimate. Based on this observation, we have conceived an algorithm that requires no more than 20 ms of initial noise.

First of all, due to the delay induced by the haircell LPF, one needs to correct the incoming $q(n)$ with the unitstep-response $w(n)$ of this filter, i.e. incoming samples $q(n)$ are replaced by $y_0 + (q(n) - y_0)/w(n)$. The maximum q_1 of the corrected samples in the first 10 ms frame (from 10 to 20 ms) is then selected as the first estimate of q_{min}. Subsequently, this estimate is refined on the basis of the averages q_j $(j = 2, 3, ...)$ of the corrected samples in the following 10 ms frames which do not exhibit speech-like characteristics:

- If $q_j < 1.3\ q_{min}$, it is assumed that the corresponding frame is still a noise frame (or speech masked by noise), and q_{min} is made equal to the average of q_1, ..., q_j.
- If $q_j > 1.3\ q_{min}$, the q_{min}-estimation algorithm is stopped, and q_{min} is left unaltered.

Note that the algorithm can stop at different values of j in the different channels of the auditory model.

Using the obtained q_{min}, the initial value of Δ can be calculated, and the actual analysis of the utterance can proceed according to the standard procedure (section 2.1).

2.3. Results

We report on experiments with white noise. Other experiments including pink noise, high frequent noise and cocktail party noise lead to similar conclusions. The phone recognizer (speaker independent continuous speech) uses a discriminative stochastic segment approach, relying on multi-layer perceptrons to estimate posterior phone probabilities [3]. The test database is a Flemish corpus containing 3753 phones. As can be seen from figure 4, nonlinear NMS is superior to linear NMS, especially for low SNRs.

3. LOUDNESS MEAN NORMALIZATION

As each channel analyses a narrow frequency band, the combined effect of convolutional noise and/or speech level variations manifests itself as an attenuation or amplification $H(n)$. Since the firing rate (f) relates to a logarithmic power ($\log P$) (see [1]), it is possible to apply mean normalization [4] in the firing rate domain. If the average loudness $\overline{f(n)}$ is measured in speech, and translated to $\overline{\log P(n)}$, one obtains that $H(n)$ emerges from

$$\overline{\log P(n)} - E[\log P] = \log H^2(n)$$

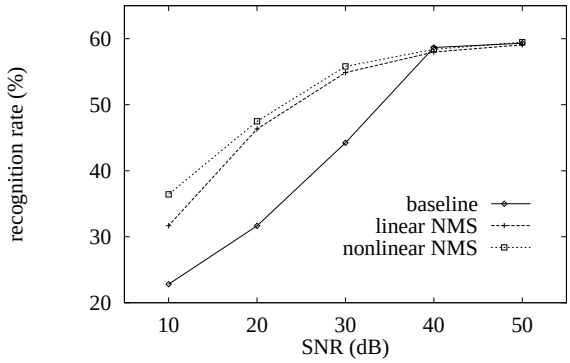

Figure 4: Phone recognition rate as a function of the global SNR. The test database is corrupted with white noise.

with $E[\log P]$ being the expected value of $\log P(n)$ in the channel, retrieved from the training database. Hence, applying mean normalization in the firing rate domain boils down to transforming $f(n)$ to $\log P(n)$, replacing $\log P(n)$ by $\log P(n) - \log H^2(n)$, and back-transforming this normalized value to a corrected firing rate. However, as argued before, we could also pursue a mean normalization before the signals are supplied to the haircell model. I.e., to counteract $H(n)$, we introduce an additional gain $G(n) = 1/H(n)$ in front of the center clipper (figure 2).

3.1. Implementation

First of all, the firing rates $f(n)$ are low-pass filtered (time constant of 36 ms) to produce loudnesses. Then these loudnesses are accumulated across channels, yielding the total loudness as a function of time. The minimal value of this loudness is measured continuously, and every 2 seconds it is used to derive a silence threshold (using an overestimation factor). The silence threshold is used to perform a speech/non-speech classification of the frames.

Every 2 seconds, $\overline{f(n)}$ in a channel is computed as the mean of the loudnesses observed in the *speech* frames found in the preceding 2 seconds. If less than half the frames were speech frames, the gain $G(n)$ is left unaltered. In the other case, a new value is computed, and a gradual transition from the former to the new value is accomplished using an exponential law with a time constant of 500 ms.

3.2. Results

Speech level variations are simulated by controlling $\mathcal{A}$ (figure 1) so as to make the maximum amplitude of each utterance $\hat{s}(n)$ equal to a given constant. Figure 5 illustrates level attenuations up to 20 dB. Clearly, if the input level drops more than 6 dB, error rates of the baseline system increase rapidly. Using LMN, an attenuation up to 25 dB does not affect the performance. In fact, LMN improves the baseline performance from 59.23% to 60.43%, indicating some speaker normalization capacities of the algorithm.

To test LMN on convolutional noise, a series of second-order filters $h(n)$ (figure 1) were implemented. By controlling the position of the poles and zeroes, different frequency responses were obtained.

Figure 6 depicts some results. Only filter 10, causing an attenuation of 30 dB in the high frequency channels, could not be completely compensated by our technique. The reason is that if $\overline{f(n)} \approx 0$, no appropriate $\overline{\log P(n)}$ can be determined.

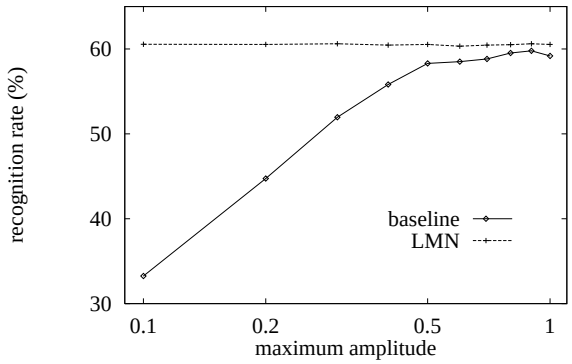

Figure 5: Recognition rate as a function of the input level.

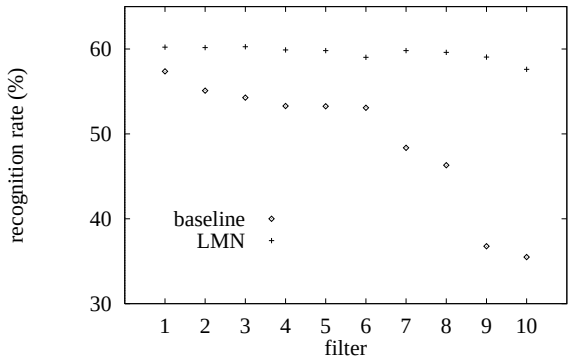

Figure 6: Recognition rates for convolutional noise. The filters are ordered by decreasing baseline performance.

4. COMBINATION OF NMS AND LMN

If the NMS-algorithm is activated, $\overline{f(n)}$ is measured on clipped noisy speech, and thus an underestimate of the average loudness of the corresponding noise free speech. Therefore, we have calculated a statistical correction so as to obtain the average log power of the noise free speech, given the average log power of the clipped noisy speech and the noise deviation $G\sigma$. I.e. if $G\sigma > y_0/0.62$, $\overline{\log P(n)}$ is replaced by $\overline{\log P(n)} + 0.57 \log(0.62 G\sigma / y_0)$.

If for some reason the NMS-algorithm is not activated, we have to make sure that the channels containing nothing but noise are not normalised. For that purpose, the minimum and maximum loudness in the channel are measured in a 2 second interval. If the difference between them (converted to log powers) is less than 15 dB, and if the average loudness is larger than its expected value (indicating that the lack of dynamics is certainly not due to a low input level), the channel is considered a noisy channel.

Tables 1 and 2 illustrate the combined effect of nonlinear NMS and LMN. Both in case of additive noise (table 1) and of a combination of different kinds of distortion (table 2), the two algorithms seem to improve the robustness of the recognizer. The results of table 2 demonstrate that even in case the individual algorithms provide but marginal improvements, the combination does extremely well.

SNR	50	40	30	20	10
NMS	40.53	41.62	44.20	52.52	63.58
NMS + LMN	39.91	41.62	43.83	51.56	64.19

Table 1: Error rates (%) in the presence of white noise.

	baseline	NMS	LMN	NMS + LMN
$r(n) + \mathcal{A}$	68.03	59.37	64.64	51.45
$r(n) + h(n)$	69.28	55.56	65.25	50.60
$h(n) + \mathcal{A}$	64.06	64.16	40.55	41.09
$r(n) + h(n) + \mathcal{A}$	63.74	63.47	64.88	50.65

Table 2: Error rates (%): $r(n)$ is white noise (20 dB), $h(n)$ is filter 8 (figure 6), and $\mathcal{A}$ corresponds to a maximum amplitude of 0.2.

5. CONCLUSION

In this paper we have introduced a nonlinear noise magnitude subtraction, which is superior to the linear noise magnitude subtraction proposed in a previous paper. To cope with convolutional noise and speech level variations, we introduced a variable gain in each channel. The gain pursues to make the average loudness in each channel independent of the input level. It was observed that the mean normalization also improves the baseline performance.

Both the noise magnitude subtraction and the loudness normalization are implemented for real-time operation, and tested separately. The combination of the two algorithms required special precautions in order to pursue that the speech rather than the speech+noise response of the channel is normalized.

6. ACKNOWLEDGEMENT

This research was performed with support of the Flemish Minister of Science Policy.

7. REFERENCES

1. Vereecken, H., and Martens, J.-P. "Recognition of noisy speech using an auditory model," *Procs EUROSPEECH*, 1995-1998, 1995.
2. Lockwood, P., and Boudy, J. "Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars," *Procs EUROSPEECH*, 79-82, 1991.
3. Martens, J.-P. "A connectionist approach to continuous speech recognition," *Procs FORWISS/CRIM ESPRIT Workshop*, 26-33, Munich, 1994.
4. Rosenberg, A.E., Lee, C.-H., and Soong, F.K. " Cepstral channel normalization techniques for HMM-based speaker verification," *Procs ICSLP*, 1835-1838, 1994.