SPECTRAL ESTIMATION AND NORMALISATION FOR ROBUST SPEECH RECOGNITION

Tom Claes, Fei Xie and Dirk Van Compernolle *

K.U.Leuven - E.S.A.T.
Kardinaal Mercierlaan 94
B-3001 Heverlee, Belgium
E-mail: Tom.Claes@esat.kuleuven.ac.be

ABSTRACT

Speech recognition in adverse conditions remains a difficult but challenging problem. It is already shown [1] that normalisation of the dynamic range (SNR[1]) of the frequency channels in a mel scale triangular filterbank (MFCC) [2], improves the robustness against both additive and convolutional noise. Nevertheless, because the method is based on a masking-technique, the improvement is small in the case of SNR values that are smaller than the target (normalised) SNR. A solution for this problem can be found in first enhancing the filterbank energies before the masking-technique is applied. For this purpose we developed a Non-linear Spectral Estimator (NSE) for speech recognition that operates on the log filterbank energies. NSE enhances these filterbank energies and makes use of SNR-normalisation also effective at very low SNRs. Experimental results are given on the NOISEX-92 [3] database. Better recognition performance is seen even at 0dB SNR.

1. INTRODUCTION

The performance of speech recognition systems drops dramatically in the presence of *additive* or *convolutional noise*, when there is no compensation for these environmental influences. In [1], we showed that the performance of a recogniser that uses MFCC-coefficients [2] as parameters, can be improved by normalising the dynamic range of the filterbank energies towards a target dynamic range (SNR). Experiments showed higher performances in both additive and convolutional noise since a better matching was achieved between the train and test environment. Nevertheless, in the case of very high noise-levels or very low SNR-values, the dynamic range can be smaller than the target dynamic range and the normalisation algorithm can not operate effectively on these signals. Figure 1 shows the masking-value as function of the measured dynamic range when the target dynamic range is set at 30dB. For a dynamic range that is smaller than the target dynamic range, the masking value is set at a fixed level of 30dB. This level doesn't have to be equal to the target dynamic range level. It is just a minimum that is always added.

In this paper we show how this problem can be solved and how the matching can be improved also in the case of low SNR-values.

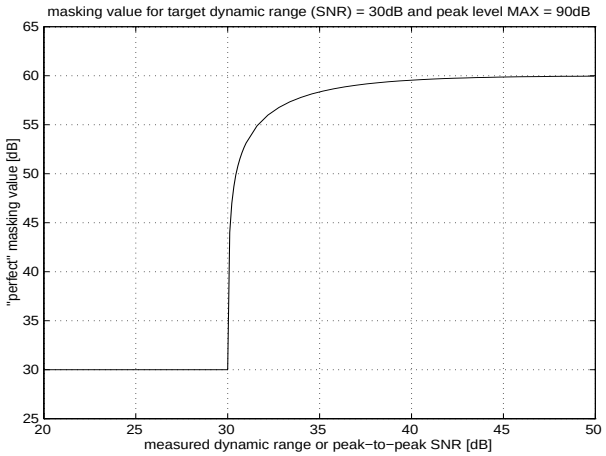

Figure 1: Masking level to get a target dynamic range or peak-to-peak SNR of 30dB for a signal with a maximum or peak level of 90dB.

2. SNR-NORMALISATION AND HIGH NOISE-LEVELS

2.1. Problem Formulation

The SNR-normalisation algorithm tries to normalise the dynamic range of the filterbank energies by adding a masking value depending on the measured dynamic range. This masking value is adapted for each new speech frame. When the measured dynamic range is larger than the target dynamic range, the masking value is increased, otherwise it is decreased. When the dynamic range of the original signal is already smaller than the target, a minimal (small) masking value is added.

Figure 2 shows the fourth frequency channel of the mel-scale filterbank of a clean and noisy signal before and after SNR-normalisation. As you can see, the matching after SNR-

*This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven, supported by Research Contract 93021 from the IWT, entitled *AKROS*.
Dirk Van Compernolle is also at Lernout and Hauspie Speech products.

[1]In this paper SNR is the peak to peak SNR or dynamic range of the filterbank energies

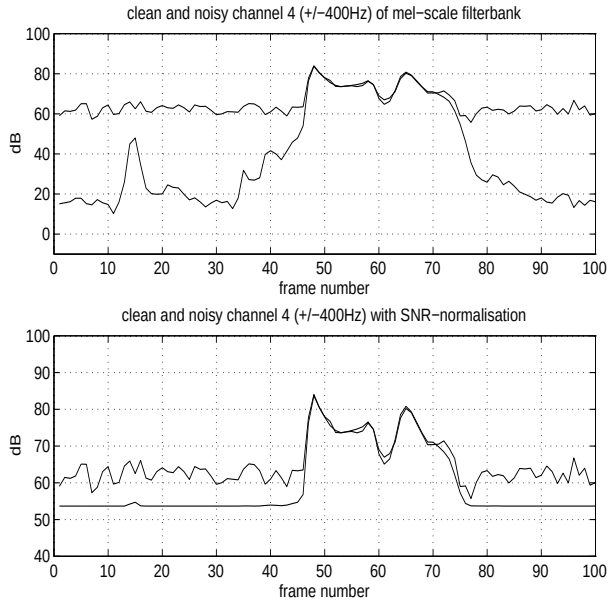

Figure 2: The plot at the top shows the fourth filterbank channel of a clean and noisy utterance 'seven'. The lower plot shows the same signals but after SNR-normalisation towards a dynamic range of 30dB. The dynamic range of the noisy signal is already smaller than 30dB and can not be normalised by adding a masking-value.

normalisation improved, but the transitions from silence (noise) to speech are still different, because there is no normalisation of the noisy signal. The reason for this is that the dynamic range of the noisy signal is already smaller than the target dynamic range and the normalisation algorithm, which is based on a masking-principle, can not operate.

We should remark that in practice the dynamic range values and also the target value are smaller than those used in the figure, but these are used for a better illustration of the problem.

2.2. SNR-Normalisation and Noise Reduction

Using noise reduction methods as Spectral Subtraction will increase the dynamic range. The use of a speech enhancement method before SNR-normalisation can in that way improve the performance of the recogniser in the case of very high noise-levels. It is also shown [4] that a controlled addition of noise (masking) after spectral subtraction improves the performance of speech recognition in noise. The reason for this is that the differences in residual noise energy after Spectral Subtraction are reduced by adding a masking-value or by normalising the dynamic range. This justifies the insertion of the 'noise-reduction' block in figure 3 for the calculation of the recognition parameters.

2.3. Non-linear Spectral Estimation

As mentioned in the previous section, noise reduction will make the SNR-normalisation technique more effective at lower SNRs. Noise reduction can either be implemented in the time domain or directly

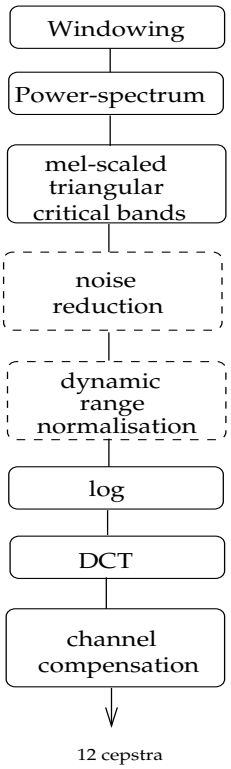

Figure 3: MFCC parameters with noise-reduction and SNR-normalisation

incorporated in the feature extraction process. In the former case, many speech enhancement techniques are readily available, such as Non-linear Spectral Estimation [5] [6] (NSE), Non-linear Spectral Subtraction [11] (NSS) and Spectral Subtraction [7] (SS). We have reported some results using GSS as preprocessing for noise reduction in [1]. However, directly incorporating the noise reduction in the feature extraction process, is more preferable for several reasons :

1. It is computational more efficient since much less frequency channels need to be processed.
2. A more accurate estimation is achieved because of the smaller variances of the log filterbank energies. [8]

In this work, we further developed our NSE to operate on the log filter-bank energies to serve as noise reduction block in the scheme depicted in figure 3. The NSE is a minimum mean square error (MMSE) estimator of the log filter-bank energies under the assumption that both speech and noise filter-bank energies have log-normal probability density functions (PDF). It should be pointed out that the filter-bank energies can be better modeled with a log-normal PDF than the power spectrum.

The NSE is formulated as:

$$\hat{s}(k) = y(k) + g(y, \mu'_y(k), \sigma_d(k))$$

where $y(k)$ is a noisy filter-bank energy, $\hat{s}(k)$ is the estimated value, $\mu'_y(k)$ is the biased speech mean after normalisation versus the noise mean and $\sigma_d(k)$ is the standard deviation of the noise. The gain function is obtained with a Mont-Carlo simulation and fitted by a trained neural network. The implementation details can be found in [6].

The upper plot of figure 4 shows the signals of the upper plot of figure 2 after NSE.

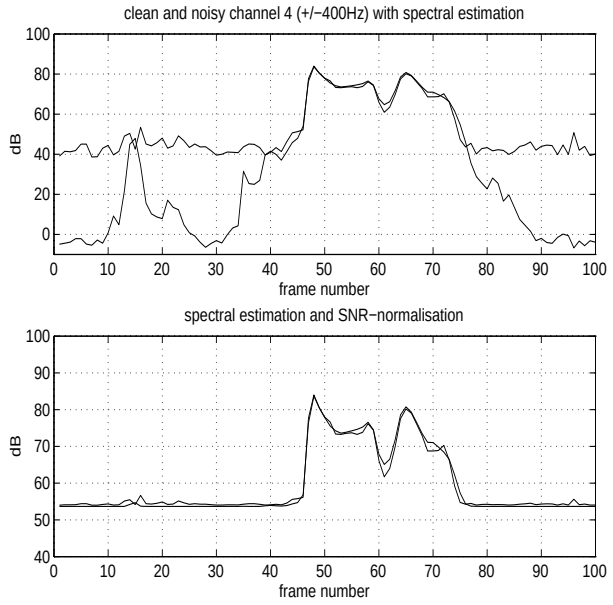

Figure 4: The plot at the top shows the fourth filterbank channel of a clean and noisy utterance 'seven' after Non-linear Spectral Estimation (NSE). The lower plot shows the same signals but after SNR-normalisation towards 30dB. Compared with figure 2 the signals are much better normalised than without NSE.

3. RECOGNITION RESULTS ON NOISEX-92

The NOISEX-92 speech-in-noise database [3] was used to evaluate the noise robustness of the presented method. The data is used at its original sampling frequency of 16kHz. It is a small vocabulary speaker dependent database. The tests are done on the *digit triplets* of the male speaker, which seemed to be the more difficult one. The experiments are done for car noise, F-16 noise and Lynx noise for global SNR ratio's from - 6 to +18 dB. and also with spectrally distorted data. The spectral tilt of NOISEX-92 is used. It has a flat response up to a break point frequency of 250 Hz followed by a +3dB/oct tilt above 250 Hz.

The MFCC features are used in combination with the noise masking technique and/or the NSE. For comparison we also used spectral subtraction (SS) [7] for the enhancement of the filterbank energies. The overestimation parameter for SS α=2.0 and the noise spectral floor β=0.01. A discrete density HMM speech recognition system with one model per word is used. For the vector quantisation 4 codebooks are used (weighted cepstra, delta-cepstra, delta^2-cepstra and delta-energy+delta^2-energy). Baseline experiments are done, i.e. training with the clean train samples, testing with noisy test samples. The recognition accuracies are given in tables 1 and 2 for car noise, F-16 noise and Lynx noise without and with spectral tilt respectively. The accuracies are calculated as $\frac{N-S-D-I}{N} * 100\%$ with S substitution errors, D deletions and I insertions for N test tokens. M12dB means that SNR-normalisation is done with a target dynamic range (SNR) of 12dB. We only give results with this level here, which was the optimal one. In [1] also results with other masking levels are given. In all cases reported in tables 1 and 2, a rasta-filtering is performed of the parameters. Without this filtering the results are worse. As you can see in the tables, the results are optimal for the combination of the SNR-normalisation technique with the Non-linear Spectral Estimator. Results of the order of 90% accuracy are achieved at 0dB SNR.

SNR (dB)	MFCC rasta	MFCC M12dB rasta	MFCC NSE rasta	MFCC SS rasta	MFCC NSE M12dB rasta	MFCC SS M12dB rasta
Car noise						
-6	18	52	31	30	63	39
0	25	66	71	57	90	64
6	47	94	82	75	99	77
12	81	99	95	88	99	87
18	92	99	99	94	100	91
F-16 noise						
-6	13	13	26	20	49	31
0	15	51	67	40	87	63
6	45	91	95	77	95	79
12	87	97	99	91	99	88
18	97	98	100	97	99	90
Lynx noise						
-6	12	15	23	18	47	25
0	15	46	60	37	89	58
6	35	91	88	76	97	81
12	85	98	98	89	99	85
18	95	99	100	95	99	89

Table 1: Summary of the test results on *digit triplets* for speech corrupted by *Car noise*, *F-16 noise* and *Lynx noise* without spectral tilt. M12dB means that SNR-normalisation is done with a target dynamic range of 12dB. SS stands for Spectral Substraction and NSE is the Non-linear Spectral Estimator.

4. CONCLUSIONS

A Non-linear Spectral Estimator (NSE) is used for the enhancement of noisy filterbank energies. These are used for calculation of MFCC coefficients [2] for speech recognition. The enhanced parameters give better recognition performance in noise. We also showed that the use of a NSE makes the use of the SNR-normalisation technique presented in [1], more effective, even with very high noise levels, where the dynamic range (SNR) is smaller than the target dynamic range. Results on NOISEX-92 [3] show accuracies of 90% on the digit triplets even at 0dB SNR.

SNR (dB)	MFCC rasta	MFCC M12dB rasta	MFCC NSE rasta	MFCC SS rasta	MFCC NSE M12dB rasta	MFCC SS M12dB rasta
Car noise						
-6	15	19	23	23	42	23
0	20	45	39	37	79	30
6	31	87	61	58	91	47
12	59	98	83	80	98	59
18	88	98	93	89	98	62
F-16 noise						
-6	13	15	18	13	23	23
0	15	31	37	33	61	51
6	23	79	79	64	93	81
12	61	95	95	85	99	89
18	93	100	99	95	99	93
Lynx noise						
-6	11	16	18	13	21	23
0	14	23	33	24	59	39
6	24	75	64	49	95	77
12	46	96	93	81	99	93
18	88	99	98	94	100	92

Table 2: Summary of the test results on *digit triplets* for speech corrupted by *Car noise*, *F-16 noise* and *Lynx noise* with spectral tilt. M12dB means that SNR-normalisation is done with a target dynamic range of 12dB. SS stands for Spectral Substraction and NSE is the Non-linear Spectral Estimator.

5. REFERENCES

1. T. Claes and D. Van Compernolle. SNR-normalisation for robust speech recognition. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, Atlanta, Georgia, U.S.A., May 7-10 1996.
2. S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Trans. on Acoustics, Speech and Signal Processing*, pages Vol. ASSP–28, No. 4 pp.357–366, 1980.
3. A. Varga, H.J.M Steenneken, M. Tomlinson and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition, 1992. Documentation included in the NOISEX-92 CD-ROMs.
4. D. Van Compernolle. Noise Adaptation in a Hidden Markov Model Speech Recognition System. *Computer Speech and Language*, 3(2):151–168, 1989.
5. F. Xie and D. Van Compernolle. A family of MLP based nonlinear spectral estimators for noise reduction. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 53–56, Adelaide, Australia, April 19-22 1994.
6. F. Xie and D. Van Compernolle . Speech Enhancement by Nonlinear Spectral Estimation - A Unifying Approach. To appear in *Speech Communication*, Vol.18, No.4 1996.
7. M. Berouti, R. Schwartz and J. Makhoul. Enhancement of speech corrupted by additive noise. In *Int. Conf. Acoust. Speech & Signal Processing*, pages 208–211, 1979.
8. A. Erell and M. Weintraub. Energy Condition Spectral Estimation for Recognition of Noisy Speech. *IEEE Transactions on Speech and Audio Processing*, 1.1:84–89, 1993.
9. J.P. Openshaw and J.S. Mason. On the limitations of cepstral features in noise. In *Int. Conf. Acoust. Speech & Signal Processing*, pages II 49–52, Adelaide, Australia, April 19-22 1994.
10. J.E. Porter and S.F. Boll. Optimal Estimators for Spectral Restoration of Noisy Speech. In *Int. Conf. Acoust. Speech & Signal Processing* , pages 18.A.2.1–4, 1984.
11. P. Lockwood and J. Boudy. Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars. *Speech Communication*, 11.2-3:215–228, 1992.
12. Y. Ephraim and D. Malah. Speech Enhancement using a Minimum Mean-Square Log-Spectral Amplitude Estimator. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-33 No.2:443–445, 1985.
13. F. Xie and D. Van Compernolle. Speech Enhancement by Nonlinear Spectral Estimation - A Unifying Approach. In *Proceedings EUROSPEECH 93*, pages 617–620, Berlin (Germany), September 21-23 1993.
14. K. Hornik, M. Stinchcombe and H. White. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2:359–366, 1989.
15. M. Trompf, R. Richter, H. Eckhardt and H. Hackbarth. Combination of Distortion-Robust Feature Extraction and Neural Noise Reduction For ASR. In *Proceedings EUROSPEECH 93*, pages 1039–1042, Berlin (Germany), September 21-23 1993.