An Assessment of Local Non-linear Spectral Subtraction for Remote Speech Recognition

Nicholas W. D. Evans and John S. Mason
Department of Electrical & Electronic Engineering,
University of Wales Swansea, SA2 8PP, UK
email: {eeevansn, J.S.D.Mason}@swansea.ac.uk

November 17, 2000

Abstract

This paper addresses the task of automatic speech recognition (ASR) in adverse conditions, in particular from within a domestic car.

Variations of non-linear spectral subtraction (NLSS) are assessed in terms of ASR relevant spectral matching in an attempt to find optimum conditions invariant across typical car noises. Lockwood [1, 2] reports the benefits of an SNR-dependent scaler in subtracting over-estimates of the noise spectrum. In this paper the idea is extended to frequency-dependent noise floors. Frequency-dependent noise floors, reflecting speech spectral distributions are proposed. With a simple linear approximation these are shown to offer small but finite improvements at no extra computational cost.

1 Introduction

In the last twenty years or so, much research has been conducted in the field of automatic speech recognition (ASR). There are ever-growing opportunities to utilise this technology in telephony generally and in particular in mobile communication.

Many companies now aim to use basic, small-to-medium vocabulary speech recognition systems to assist in telephony services. Typical environments where this technology might be used are at home, in the office, in the street or in the car, in other words in a wide range of everyday situations. Inevitably therefore, a wide range of background noises and levels of noise will be encountered. The influence on ASR performance of such additive noise is well documented and there has been a vast amount of research effort in the area of robust ASR. Recent surveys include [3, 4, 5].

Approaches to reduce the effects of speech degradation tend to be geared to the class of noise and to the application itself. The goal of speech enhancement is generally interpreted as the reduction of noise to improve audible quality. The term noise compensation on the other hand is often taken to mean the reduction of the effects of noise to improve the robustness of ASR systems. These two interpretations are used below. One important difference between the two stems from the intended receiver of the processed signal. A signal is likely to be processed differently for a human receiver than for a machine. In the latter case, where recognition performance is the sole criterion, the noise compensation should be matched to the recognition process, and audible speech quality as perceived by the human is irrelevant.

This paper focuses on in-car mobile communication and its associated background additive noise. The signal is telephony band-limited and noise compensation is performed locally prior to communication channel processing.

Non-linear spectral subtraction (NLSS) has been applied both speech enhancement [6, 7] and noise compensation [1, 2, 8, 9] and uses a spectral estimation of the noise obtained during speech pauses. This is then used to compensate a speech signal degraded by additive noise. Here the experimental study aims to quantify the effectiveness of various NLSS techniques reported in the literature. Short-term spectral estimates of typical in-car noise are examined and quantitative results for NLSS are given.

Most variations of NLSS use an SNR-dependent over-estimation of the noise in the subtraction process. We extend this idea to frequency-dependent noise floors. The hypothesis is that noise floors can be beneficial not only in reducing musical noise but also in preventing speech loss particularly when adopting the popular practice of subtracting over-estimates of the noise.

In considering assessment strategies it is important to consider the application. Here the end-goal is ASR and many previous publications have used ASR error rates as the cost function to minimise. This ties the results very much to individual ASR characteristics and performance levels. One step removed from a given ASR system is spectral matching, and since essentially all ASR systems use spectral matching, and the large majority a mel-scaled form of spectral matching, it is this latter form that is adopted here.

2 ASR Performance in Noise

2.1 Additive Noise

It is well known that when the training and testing environments of conventional ASR systems differ, performance tends to be impaired. Such a situation is referred to as the *mis-match* between training and testing conditions. These differences may be as subtle as a different microphone, the presence of other speakers referred to as the *cocktail party effect*, or the acoustic effects of the immediate environment such as echo or background noise.

In the car situation, it can often be assumed that for a short period of time, that is for a few seconds, the noise characteristics remain constant. It is this consistency in the noise that is exploited in spectral subtraction. A noise estimate is subtracted from a contaminated signal, which yields an approximation of the original, undegraded clean speech.

2.2 Lombard Reflex

The effects of degradation imposed by additive noise are relatively well understood in the context of ASR. In contrast the effects of the Lombard reflex on a speaker or more importantly their speech, are far less well understood and often overlooked, Junqua [10].

When a speaker is subjected to noise they tend to modify their speech as a reflex to the condition. The intention is to maintain intelligible communication [11, 12]. Changes in vocal effort, intelligibility, word duration, vocabulary, phonetic information and formant energy are all reported in the literature [13, 14, 15, 16]. Junqua [10] gives a good survey of Lombard related work.

In the current study it was observed that much of the noise within a car can be outside of the telephony spectrum. If such out-of-band noise is the only noise present then the transmitted signal would be noise-free. Thus while the speech would be relatively clean it nonetheless could well be subjected to the Lombard reflex.

Additionally, the extent of the Lombard effect is heavily speaker dependent [10] which accounts for variation in performance of ASR systems particularly speaker-dependent systems without suitably broad training.

It is the variation in speech characteristics due to the speaker or the environment that present difficulties for ASR systems trained on speech not adequately reflecting these variations. Adapting speech models to take into account Lombard effects has proved to be a very difficult task [10]. The most likely successful compensation for the Lombard effect is matched speaker-specific training. Such is the effect of additive noise and Lombard speech on ASR systems, that one recent initiative [17] has sought to obtain speech models produced in the car for several European languages. Systems trained on this sort of speech will inevitably be more robust to the effects of Lombard speech.

2.3 Coding and Lossy Compression

The position of a speech compensation process in the communications system is likely to have a critical influence on its performance. Figure 1 illustrates a typical GSM communications process. At the near end (in the car in our scenario), the signal is first digitized and analysed to give source model parameters followed by a non-linear lossy compression stage. At the far end reverse processes regenerate the speech

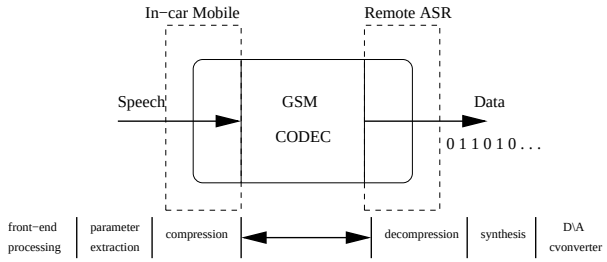

Figure 1: A typical GSM communication channel scenario

waveform. Note if the recipient is an ASR system then the synthesis and DAC stage may be avoided, though coding losses will still apply.

Noise estimates are inherently less accurate at the far end due to the lossy, non-linear nature of the coding schemes in modern telephony systems. It is for this reason that it is beneficial to perform noise compensation at the near end.

3 Non-linear Spectral Subtraction

Spectral subtraction as initially proposed by Boll [6] in 1979 is a popular algorithmic, cook-book approach of noise compensation and speech enhancement. Standard spectral subtraction relies on the relative stationarity of background noise. Based on this assumption, the noise during speech gaps is deemed representative of that during immediately following speech segments. Thus, a spectral estimate of the contaminating noise is obtained prior to speech activity and then used in the subtraction process.

Typically, this noise estimate is measured over M frames by:

$$\begin{aligned} \overline{|N(e^{j\omega})|} &= \frac{1}{M}\sum_{i=1}^{M-1}|N(e^{j\omega})| \\ M &= \text{number of frames} \end{aligned} \tag{1}$$

where $N(e^{j\omega})$ is the instantaneous noise estimate and $\overline{|N(e^{j\omega})|}$ is the average estimate used in NLSS.

There are many variations on how the noise estimate is used. The most popular is that where an over-estimate of the spectral noise power is subtracted from the degraded signal to suppress what Berouti et al [7] termed *musical noise*.

Musical noise occurs when the processed signal contains large peaks in the spectrum relating to noise which was unsuccessfully removed. These spectral peaks manifest themselves in the time domain as random tonal bursts. To improve noise reduction and suppress musical noise Berouti introduced two new parameters to the standard spectral subtraction process: α a factor controlling the amount of noise subtracted, and β, a noise floor corresponding to a fraction of the original degraded signal. Berouti's approach is occasionally referred to in the literature as generalised spectral subtraction.

$$\begin{aligned} \text{Let } P(\omega) &= P_D(\omega) - \alpha P_N(\omega) \\ \hat{P}_S(\omega) &= \begin{cases} P(\omega), \text{ if } P(\omega) > \beta P_N(\omega) \\ \beta P_N(\omega), \text{ otherwise} \end{cases} \end{aligned} \tag{2}$$

where $P_D(\omega)$, $P_N(\omega)$ and $\hat{P}_S(\omega)$ are the original degraded, noise estimate and estimated speech spectra respectively. Using such an approach it has been reported that non-physical negative values in the processed spectrum are removed, and the presence of musical noise is reduced [7].

For values of α>1, the energy in the noise peaks will be lower than for standard spectral subtraction. The noise spectral floor, β is introduced to further reduce spectral excursions corresponding to remaining noise peaks. The context of Berouti's original work is speech enhancement. Improved quality and minimal intelligibility loss in processed speech is the goal.

Berouti's algorithm has been adopted by many. Lockwood et al [1, 2] employ a similar, non-linear implementation but apply it to noise compensation in an ASR context. Lockwood employs a frequency-dependent, over-estimation of the noise, $\alpha(\omega)$ where the maximum of the local noise mean is used to calculate optimum values for $\alpha(\omega)$, opt$(\alpha(\omega))$. In this manner a higher subtraction factor is applied in low SNR regions than for high SNR regions where an over-estimate of the noise may cause increased speech distortion. Lockwood reports optimum results where $1 \leq$ opt$(\alpha(\omega)) \leq 3$.

Schless and Class [9] also adopted Berouti's approach but introduced SNR-dependent over-estimation and SNR-dependent noise flooring. Spectral subtraction is successfully combined with another technique of robust ASR, parallel model combination (PMC) [18, 19] which relies on models of undistorted clean speech. The motivation of the study is to minimise distortion introduced to a signal processed by spectral subtraction so that the PMC stage need not be modified. In effect, the approach attempts to compensate for the mis-match in the estimated and instantaneous noise spectra. With a combination of spectral subtraction and PMC, optimum results are obtained with $\alpha \leq 1.0$ and $\beta \geq 0.15$. Their precise values are set according to a linear function, dependent on SNR.

In another example [18], spectral subtraction is performed after the application of a mel-filter bank but better results are reported by applying the compensation before mel smoothing, hence this is the approach adopted here.

4 Assessment

Using spectral subtraction in an extreme manner, it is of course possible to remove all the noise only with severe degradation of any speech signal present. There is an inherent trade-off between noise reduction and speech distortion. In this context our goal is to maximise noise reduction and minimise speech distortion. An appropriate measure therefore, is spectral matching.

This approach to assessment is adopted because it mirrors the similarity measures central to nearly all ASR systems. Using such an assessment technique, it becomes a relatively simple task to alter the parameters used for spectral matching to those used in a particular ASR system. Thus quantitative assessment of NLSS methods are presented in terms of *mel-scaled* spectral distance measures.

Fourteenth order mel-cepstra are generated for both the original clean speech and for the NLSS processed utterances. The normalised difference, S_{diff}, between the two sets of spectra are calculated over time T and k frames by:

$$S_{diff} = \frac{1}{kN} \frac{\sum^{T} (X^N - Y^N)^2}{\sum^{T} (X^N)^2} \tag{3}$$

where Y^N is the processed spectra and X^N is the clean spectra computed over N samples which acts to normalise the measure.

5 Experimental Work

5.1 Noise Estimation

A suitable period for noise estimation is essential if an accurate approximation of the instantaneous noise spectrum is to be found. From a database of sample sounds recorded in a domestic car, an experiment was devised to find such a suitable value. A number of successive difference frames obtained from sequential mean noise frames as in Equation 1 are averaged to give $\overline{D[i]}$ where $i = 0..N$ and $N = 256$ samples per frame of a noise signal sampled at 8kHz. The normalised squared difference, $\tilde{D}$ is obtained by:

$$\tilde{D} = \frac{1}{T} \sum_{i}^{T} \overline{D[i]} \tag{4}$$

A plot of the power in the normalised squared difference, $\tilde{D}$ against the measurement interval, T is presented in Figure 2. The five profiles relate to different road speeds, 0, 10, 35, 50 and 70mph. It is

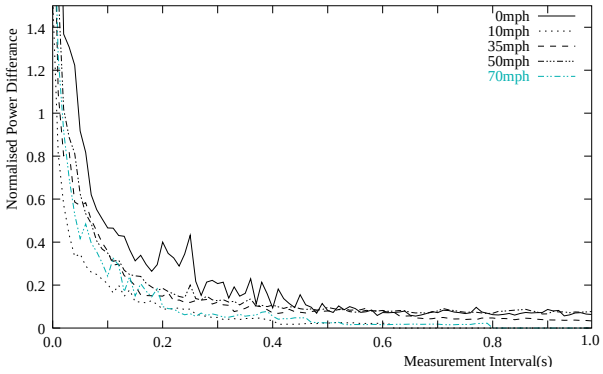

Figure 2: The power in the normalised difference, $\tilde{D}$, against the measurement interval, T, for five different car speeds

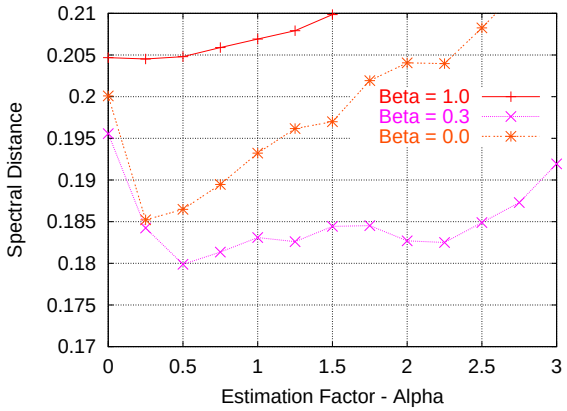

Figure 3: Using the noise estimation to calculate the noise floor

observed that no significant reduction in the normalised squared difference was obtained for values of T over approximately 0.5 seconds, for all speeds.

5.2 Non-linear Spectral Subtraction

Using an over-long period (1 second) to obtain noise estimates, power subtraction and half-wave rectification, experiments were conducted to assess the performance of various non-linear spectral subtraction schemes. The sampling frequency was fixed at 8kHz in line with GSM mobile telephones. The frame size and frame rate where fixed at 32ms and 0.125ms (one sample) for optimum estimation. Figure 3 illustrates the results of preliminary experiments where the noise estimation factor, α is varied with the spectral floor, β as in Equation 2. For each profile a non-SNR-dependent value of α is used. The spectral distance is plotted against the noise estimation factor, α. It can be seen that for $\beta = 0.3$, noise estimation factors of $0.5 \leq \alpha \leq 2.5$ yield the most improved spectral distance measures than without a spectral floor, $\beta = 0.0$. Values of $\beta \geq 0.3$ tend to retain excess noise. Larger values of α distort the signal and the spectral distance measures deteriorate as for the profile of $\beta = 1.0$.

Schless and Class [9] used an alternative approach to spectral noise flooring than that used by Berouti where the noise floor is determined from the degraded signal rather than from the noise estimate:

$$\begin{aligned} \text{Let } P(\omega) &= P_D(\omega) - \alpha P_N(\omega) \\ \hat{P}_S(\omega) &= \begin{cases} P(\omega), \text{ if } P(\omega) > \beta P_D(\omega) \\ \beta P_D(\omega), \text{ otherwise} \end{cases} \end{aligned} \tag{5}$$

Note $P_D(\omega)$ has replaced $P_N(\omega)$ in Equation 2. No explanation for this arrangement is offered in [9]. Berouti originally introduced the spectral floor to suppress musical noise, which is a likely motivation for using the noise estimate to calculate its value. However, where ASR is the context an alternative choice is the degraded signal, $P_D(\omega)$ since it may be thought of as a method of compensation for the error in the noise spectral estimate. Thus, the noise floor would compensate for speech distortion introduced by spectral subtraction, especially with favorable SNRs where $P_D(\omega)$ is likely to be more correlated with the clean speech, $P_S(\omega)$, than the noise estimate, $P_N(\omega)$.

Figure 4 illustrates results where the subtraction factor is varied as above but with a noise floor obtained from the original degraded signal as in Equation 5. A profile corresponding to zero noise floor is plotted as a baseline and is essentially the spectral distance of the original degraded signal, $P_D(\omega)$. With the new method, higher values of the noise estimation factor, $1.5 \leq \alpha \leq 2.5$ yield the better results and an improvement in spectral distance where $\beta = 0.5$.

Schless and Class [9] used an SNR-dependent noise floor in a similar fashion to the SNR-dependent estimation factor used by Lockwood [1, 2]. The function used to determine the noise floor is described as linear. If a noise floor is applied to noise compensation in order to account for spectral noise estimation

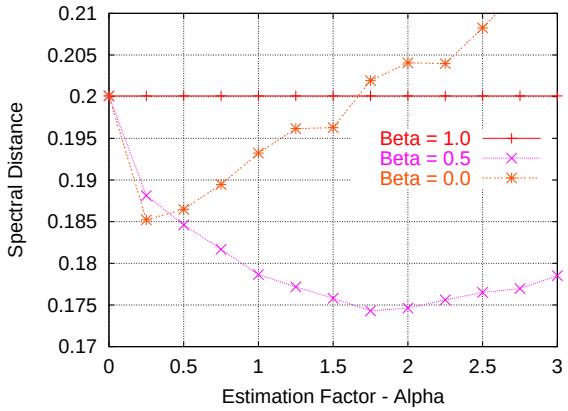

Figure 4: Using the degraded signal to calculate the noise floor

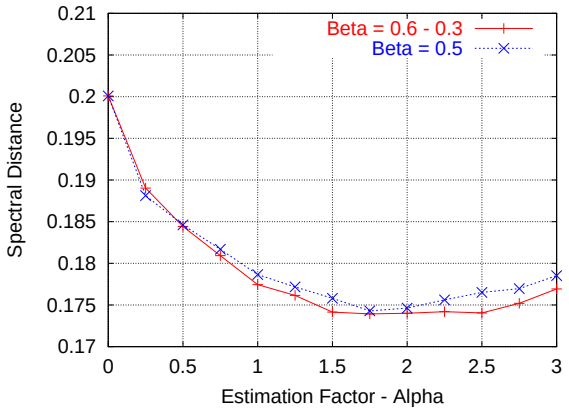

Figure 5: Using the degraded signal to calculate a non-linear noise floor

error and hence lessen speech distortion, then a non-linear spectral floor reflecting speech spectral distributions may yield better results.

Figure 5 illustrates results of experiments using a non-linear spectral floor where β is replaced by $\beta(\omega)$ in Equation 5. $\beta(\omega)$ is a simple linear approximation to the speech spectral distribution. A profile for results of NLSS using a noise floor where $0.3 \leq \beta \leq 0.6$ yields better results than for the linear noise floors in Figures 3 and 4. Note that the spectral distance is lower than for $\beta = 0.5$ and the trough is broader. Optimal values of β are still between 1.5 and 2.5.

6 Interpretation of Results

The first observation coming from the study is that any speech distortion caused by over-estimation of the noise results in a poor spectral distance measure as in Figure 3 for $\beta = 1.0$. This is understandable since the measure is calculated against distortion free, clean speech. Some noise over-estimation is inevitable since the noise is never completely stationary. It is this non-stationarity which gives rise indirectly to musical noise.

Secondly, if the noise estimate is a significant over-estimate, more noise (and speech) is removed from the signal and the effects of musical noise become more noticeable and from audible assessment, quite disturbing for the human listener. Another observation is that using the standard method of spectral subtraction an under-estimate of the noise must be used to minimise speech distortion. Thus, less noise is removed than for setups where the actual noise estimate is used. There is an obvious trade-off between speech distortion and noise reduction.

Using the modified approach to spectral subtraction as in Equation 5 and Figure 4, improved results are obtained. For similar spectral distance measures, an over-estimate of the noise can be subtracted with a slightly elevated spectral floor. A likely explanation for this observation is that the spectral floor compensates for the inaccuracy in the spectral noise estimation since it is more correlated with the clean speech than the noise estimate as used in the standard approach.

For the non-linear implementation of the noise floor small, finite improvements in the spectral distance measure are gained. This improvement may be attributed to the correlation of the noise floor with the speech and the noise floor function aimed to reflect the speech spectral distribution.

7 Conclusions

Much of the in-car noise is outside of the telephony band thus is easily removed and will not be present in the coded signal. While the noise itself will not degrade ASR performance, the induced Lombard Reflex is likely to do so.

Approximately 0.5 seconds is sufficient to obtain a good estimation of the noise spectrum and this is consistent across different car speeds.

At best, an improvement of approximately 15% in spectral distance is obtained with NLSS. A new simple non-linear spectral noise floor reflecting the spectral distribution of speech has been introduced and shown to be beneficial. It introduces no additional computational cost.

Acknowledgements

The authors would like to thank Orange plc (St. James Court, Great Park Road, Almondsbury Park, Bristol, BS32 4QJ, UK) for their financial support and Andrew Thomas for his practical contributions throughout the work. Nicholas Evans is an EPSRC funded research student.

References

[1] P. Lockwood, C. Baillargeat, J. M. Gillot, J. Boudy, and G. Faucon. Noise reduction for speech enhancement in cars: non-linear spectral subtraction / Kalman filtering. In *Proc. EuroSpeech*, pages 83–86, September 1991.

[2] P.Lockwood and J.Boudy. Experiments with a non-linear spectral subtractor (NSS), hidden markov models and the projection, for robust speech recognition in cars. In *Proc. Eurospeech*, volume 1, pages 79–82, September 91.

[3] Y. Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16:261–291, 1995.

[4] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* Kluwer Academic Publishers, 1992.

[5] Jean-Claude Junqua and Jean-Paul Haton. *Robustness in Automatic Speech Recognition: fundamentals and applications.* Kluwer Academic Publishers, 1996.

[6] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on ASSP*, pages 113–120, 1979.

[7] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. ICASSP*, pages 212–215, April 1979.

[8] C. Mokbel, L. Mauuary, L Karray, D. Jouvet, J. Monné, J. Simonin, and K. Bartkova. Towards improving ASR robustness for PSN and GSM telephone applications. *Speech Communication*, 23:141–159, October 1997.

[9] Volker Schless and Fritz Class. SNR-Dependent Flooring and Noise Overestimation for Joint Application of Spectral Subtraction and Model Combination. In *Proc. ICSLP*, 1998.

[10] Jean-Claude Junqua. The Lombard Reflex and its role on human listeners and automatic speech recognizers. *J. Acoust. Soc. Am.*, January 1993.

[11] J.J. Eagan. "Psychoacoustics of the Lombard Voice Reflex", Ph.D. Thesis, Western Reserve University, 1967.

[12] E. Halphen. "Des Lésions Traumatiques de l'Oreille Interne", Thèse de la Faculté de Médecine Paris, 1910.

[13] J.M. Pickett. Effects of Vocal Force on the Intelligibility of Speech Sounds. *J. Acoust. Soc. Am.*, 28(5):902–905, 1956.

[14] D. Howes. On the Relation between the Intelligibility and Frequency of Occurrence of English Words. *J. Acoust. Soc. Am.*, 29:296–305, 1957.

[15] D. Rostolland and C. Parant. Distortion and Intelligibility of Shouted Voice. In *In Symposium: Speech Intelligibility. Liége*, pages 293–304, 1973.

[16] B. J. Stanton, L. H. Jamieson, and G. Allen. Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions. In *Proc. ICASSP*, pages 331–334, 1988.

[17] Asunción Moreno, Børge Lindberg, Christoph Draxler, Gaël Richard, Khalid Choukri, Stephan Euler, and Jeffrey Allen. SpeechDat-CAR. A Large Speech Database for Automotive Environments. In *Proc. LREC, Athens*, volume 2, pages 895–900, 2000.

[18] J.A. Nolazco Flores and S.J. Young. Continuous Speech Recognition in Noise using Spectral Subtraction and HMM Adaptation. In *Proc. ICASSP*, volume 1, pages 409–412, 1994.

[19] M. Gales and S. Young. Robust Continuous Speech Recognition using Parallel Model Combination. *IEEE Trans. on Speech and Audio Processing*, 4(5):352–359, 1996.