AN ASSESSMENT ON THE FUNDAMENTAL LIMITATIONS OF SPECTRAL SUBTRACTION

Nicholas W. D. Evans, John S. D. Mason, Wei M. Liu and Benoît Fauve

School of Engineering, University of Wales Swansea,
Singleton Park, Swansea, SA2 8PP, UK

email: {n.w.d.evans, j.s.d.mason, 199997, 191992}@swansea.ac.uk

ABSTRACT

As with many approaches to noise robust automatic speech recognition (ASR) the benefits of spectral subtraction tend to diminish as noise levels in the order of 0 dB are approached. Whilst the majority of related work focuses on reducing magnitude errors a number of new approaches addressing the often overlooked, additional sources of error have appeared in the literature in recent years. Relatively lacking in the literature, however, is an empirical assessment which compares the effects of each error when noisy speech is processed by spectral subtraction. Such studies are vital in order to appreciate the potential penalty in performance when sources of error are overlooked.

The objective in this paper is to assess, through ASR, the performance penalty associated with each source of error when noisy speech is treated with spectral subtraction. Experimental evidence based on two standard European databases and ASR protocols illustrates that, perhaps contrary to popular belief, for noise levels in the order of 0 dB and below, these often overlooked sources of error can lead to non-negligible degradations in performance. Whilst not a new idea, here the original emphasis is a thorough assessment that empirically highlights both the fundamental limitations and potential benefit of including the full complement of errors in the spectral subtraction model.

1. INTRODUCTION

Proposed by Boll in 1979 [1] spectral subtraction is one of the earliest and longest standing approaches to noise compensation and speech enhancement. As in the subsequent work of Lockwood *et al* in 1991 [2], and this paper, the majority of recent literature with a spectral subtraction theme focuses on noise robust automatic speech recognition (ASR). A literature search reveals an abundance of related research papers, both long past and recent. However, despite the many thousands of man hours spent optimising spectral subtraction, improvements in ASR performance tend to diminish as noise levels in the order of 0 dB are approached. In fact, it is difficult to find publications that report any improvements in *intelligibility* through the processing of speech by spectral subtraction.

The early work of Lim in 1979 [3] provides a theoretical analysis of the error sources associated with spectral subtraction, namely phase, cross-term and magnitude errors. Whilst the majority of related work focuses on reducing magnitude errors a number of new approaches addressing the additional error sources have appeared in the literature in recent years. In 2004, noting that even given accurate noise statistics the phase relationship between speech and noise can lead to negative values in the processed spectrum, Xu *et al* [4] showed that full-wave rectification better retains speech information and provides optimum orthogonality between the estimated noise and speech signals. Although aimed at feature enhancement as opposed to speech enhancement, Deng *et al* [5] proposed a new probabilistic, nonlinear acoustic environment model to incorporate the aforementioned phase relationship. A quantitative analysis of how noise affects the speech spectrum is provided by Zhu and Alwan [6].

Relatively lacking in the literature, however, is an empirical assessment which compares the effects of each error when noisy speech is processed by spectral subtraction. Such studies are vital in order to appreciate the potential penalty in performance when cross-term and phase errors are neglected. Thus, building on our recently published work [7] and providing a new interpretation, herein lie the objectives of this paper, namely to assess, in terms of ASR word accuracy, the performance of spectral subtraction with and without each error source and hence to demonstrate the fundamental limitations.

The remainder of this paper is organised as follows. Section 2 describes a conventional implementation of spectral subtraction. The fundamental limitations that are assessed in this paper are introduced in Section 3. Section 4 describes the experimental setup. Results and discussion are presented in Sections 5 and 6 and conclusions follow in Section 7.

2. SPECTRAL SUBTRACTION

Spectral subtraction is not a recent approach to noise compensation and was proposed by Boll in 1979 [1]. There is, however, much more recent work in the literature relating to different implementations and configurations of spectral subtraction. Thus the objective here is to describe what is perhaps best termed as a *conventional* implementation of spectral subtraction drawing from [1, 8].

The goal of spectral subtraction is the suppression of additive noise from a corrupt signal. Speech degraded by additive noise can be represented by:

$$d(t) = s(t) + n(t), \tag{1}$$

where $d(t)$, $s(t)$ and $n(t)$ are the *d*egraded or corrupt speech, original clean *s*peech (no added noise) and *n*oise signals respectively. From the discrete Fourier transform (DFT) of sliding frames typically in the order of 20-40 ms, an estimate of the original clean speech is obtained in the frequency domain by subtracting the noise estimate from the corrupt power spectrum:

$$|\hat{S}(e^{j\omega})|^2 = |D(e^{j\omega})|^2 - |\hat{N}(e^{j\omega})|^2, \tag{2}$$

where the $\hat{}$ symbol indicates an estimate as opposed to observed signals. The assumption is thus made that noise reduction is achieved by suppressing the effect of noise from the magnitude spectra only. The subtraction process can be in power terms as in Equation 2 or in true magnitude terms, i.e. using the square roots of the terms in Equation 2.

For speech enhancement applications, where a time domain representation is sought, a complex estimate (magnitude and phase), $\hat{S}(e^{j\omega})$, is required and in practice this is obtained by combining the enhanced magnitude with the phase of the corrupt spectrum, $\theta_D(e^{j\omega})$:

$$\hat{S}(e^{j\omega}) = \left[\ |D(e^{j\omega})|^2 - |\hat{N}(e^{j\omega})|^2\ \right]^{1/2} e^{\theta_D(e^{j\omega})} \quad (3)$$

A time domain representation is then resynthesised via the inverse DFT. Negative values at any frequency, ω, occur whenever $|\hat{N}(e^{j\omega})| > |D(e^{j\omega})|$ and thus generally necessitate some form of post-processing prior to resynthesis since they have no physical meaning.

Nearly all later work advocates noise over-estimates and noise floors, as introduced by the early original work of Berouti *et al* [8]. Equation 3 is thus modified to:

$$\begin{aligned}\hat{S}(e^{j\omega}) \quad = \quad & \max\left(\left[\ |D(e^{j\omega})|^2 - \alpha|\hat{N}(e^{j\omega})|^2\ \right],\right. \\ & \left.\beta|D(e^{j\omega})|^2\ \right)^{1/2} e^{\theta_D(e^{j\omega})}, \end{aligned} \quad (4)$$

where α is the noise over-estimation parameter and β is the noise floor as in [2, 8]. The idea is to increase noise attenuation through α and to suppress musical noise and negative values in the processed magnitude spectrum through β. Both α and β are tunable parameters and are optimised according to the noise level.

3. FUNDAMENTAL LIMITATIONS

The objective in this section is to introduce the sources of error in a conventional implementation of spectral subtraction. In [3] it is shown that the clean speech spectrum, $S(e^{j\omega})$, in exact terms, is expressed by:

$$\begin{aligned}S(e^{j\omega}) \quad = \quad & \left[\ |D(e^{j\omega})|^2 - |N(e^{j\omega})|^2 - S(e^{j\omega}) \cdot N^*(e^{j\omega})\right. \\ & \left. - S^*(e^{j\omega}) \cdot N(e^{j\omega})\ \right]^{1/2} e^{\theta_S(e^{j\omega})}, \end{aligned} \quad (5)$$

where * denotes the complex conjugate. $S(e^{j\omega}) \cdot N^*(e^{j\omega})$ and $S^*(e^{j\omega}) \cdot N(e^{j\omega})$, termed throughout this paper as cross-terms to reflect the above notation, may alternatively be reduced to $2 \cdot |S(e^{j\omega})| \cdot |N(e^{j\omega})| \cdot \cos\theta$ where θ represents the phase difference between $S(e^{j\omega})$ and $N(e^{j\omega})$.

The three sources of error inherent in a conventional implementation of spectral subtraction are thus evident upon the comparison of Equations 3 and 5. They are phase, cross-term and magnitude errors.

3.1. Phase Errors

Phase errors do not ordinarily affect ASR performance when features are derived from the same short term magnitude spectra used in the spectral subtraction process. However, returning to a time domain representation of the processed speech will introduce phase errors associated with the differences between $\theta_S(e^{j\omega})$ and $\theta_D(e^{j\omega})$ and will influence subsequent feature extraction. Phase errors are considered in this paper to embrace situations where it is desirable to produce an enhanced time domain speech signal as an intermediary stage before ASR, i.e. where the enhanced speech data is communicated rather than their features as in the ETSI Aurora 2 *distributed* ASR ethos. Utilising noise over-estimates and noise floors, the effect of phase errors may be assessed by modifying Equation 5 to:

$$\begin{aligned}\hat{S}(e^{j\omega}) = \max\left(\left[|D(e^{j\omega})|^2 - \alpha|N(e^{j\omega})|^2 - S(e^{j\omega}) \cdot N^*(e^{j\omega})\right.\right. \\ \left.\left. -\ S^*(e^{j\omega}) \cdot N(e^{j\omega})\ \right], \beta|D(e^{j\omega})|^2\ \right)^{1/2} e^{\theta_D(e^{j\omega})} \end{aligned} \quad (6)$$

3.2. Cross-term Errors

Cross-term errors are also commonly assumed to have a negligible influence since, in the expected sense, $S(e^{j\omega}) \cdot N^*(e^{j\omega})$ and $S^*(e^{j\omega}) \cdot N(e^{j\omega})$, average to zero and are generally omitted. To include cross-term errors only and resynthesising with the phase of the original speech, $\theta_S(e^{j\omega})$ (i.e. no phase errors), the subtraction is then implemented as:

$$\begin{aligned}\hat{S}(e^{j\omega}) = \max\left(\left[\ |D(e^{j\omega})|^2 - \alpha|N(e^{j\omega})|^2\ \right],\right. \\ \left.\beta|D(e^{j\omega})|^2\ \right)^{1/2} e^{\theta_S(e^{j\omega})} \end{aligned} \quad (7)$$

Note that even though cross-terms are related to the *phase* difference between the clean speech and noise they affect the *magnitude* in Equation 5.

3.3. Magnitude Errors

The third and final source of errors lies in the differences between $|N(e^{j\omega})|$ and $|\hat{N}(e^{j\omega})|$ and are referred to as magnitude errors throughout this paper. Of course cross-term errors also constitute magnitude errors but two separate definitions are adopted to reflect the general assumption that cross-term errors are negligible, i.e. in practice the procedure focuses only on obtaining effective estimates of $|N(e^{j\omega})|$.

The contribution of magnitude errors may be assessed by comparing the performance of spectral subtraction with known values of $|N(e^{j\omega})|$ to the performance with estimated values, $|\hat{N}(e^{j\omega})|$. The two spectral subtraction equations are then given by Equation 7, except with $e^{\theta_S(e^{j\omega})}$ replaced by $e^{\theta_D(e^{j\omega})}$, and by Equation 4. Note that in both cases phase and cross-term errors are present, the only differences are then between $|N(e^{j\omega})|$ and $|\hat{N}(e^{j\omega})|$.

4. EXPERIMENTAL SETUP

The objective of this paper is to compare, through ASR, the impact of phase, cross-term and magnitude errors when noisy speech is processed with spectral subtraction. First though, the ASR databases and recogniser configuration are described.

4.1. Databases

The ASR experiments reported in this paper were performed on the standard ETSI Aurora 2 database [9] and the Welsh SpeechDat(II) FDB-2000 database [10], hereafter referred to as the WSD(II) database. The Aurora 2 database comprises digit strings spoken by American English speakers. The car noise subset of the Aurora 2 database was selected and is justified by the popular application of spectral subtraction to in-car, noise robust ASR. The standard specifies 8440 training utterances and 1001 test utterances for each noise level. Full details of the Aurora 2 database can be found in [9].

The WSD(II) database comprises isolated digits recorded from members of the Welsh speaking public. Car noise from the Aurora 2 database was used to create a car noise subset of the WSD(II) database. There are 100 speaker training utterances and 1500 speaker test utterances. Full details of the configuration of the WSD(II) database can be found in [7, 10].

According to the Aurora 2 standard there are six different noise levels ranging from SNRs between +20 dB and -5 dB in addition to a seventh clean (no added noise) condition for both databases.

4.2. Spectral Subtraction Implementation

Spectral subtraction is adopted as a pre-processing, speech enhancement stage and a time domain representation of the enhanced signal is resynthesised prior to feature extraction. All improvements in ASR performance may therefore be attributed to spectral subtraction and not to any modifications to either the feature extraction or recognition stages. The clean speech and noise data were obtained by subtracting the noisy files from the original clean speech files in the time domain, sample by sample. The complex, frequency domain representations of the clean speech, $S(e^{j\omega})$, corrupt speech, $D(e^{j\omega})$, and corresponding noise, $N(e^{j\omega})$, were then all derived using the discrete Fourier transform (DFT) from frames of 32 ms with an overlap of 16 ms. The phase of both the degraded and original speech as well as the cross-terms in Equation 5 are then all available, thus the contribution to the degradation in spectral subtraction performance due to each error may be assessed. In all cases the noise over-estimate, α, and noise floor, β, in Equations 4, 6 and 7 are empirically optimised for each condition separately.

The noise estimate in Equation 2 is conventionally obtained during non-speech intervals and in the frequency domain from short term magnitude spectra. This approach is adopted for the experiments on the WSD(II) database and utilises 0.5 s hand-labelled non-speech intervals. However, the Aurora 2 database is not hand-labelled and so an alternative approach to noise estimation is required. Due to well known difficulties associated with accurate end-point detection, particularly under high noise conditions, quantile-based noise estimation (QBNE) [11] was used for all experiments on the Aurora 2 database. The main advantage of the quantile-based approach is that explicit speech, non-speech detection is not required. Instead, speech and non-speech regions are detected implicitly and the estimation process is continuous, spanning 0.75 s windows in both speech and non-speech intervals. In [12] the Authors' previously published work assessed QBNE on the same database demonstrating very favourable results.

4.3. ASR Configuration

In order to follow wholly standard experimental protocols the Aurora 2 W1007 standard feature extractor and HTK reference recogniser were used for ASR experiments, the full details of which can be found in [9]. In summary, 39^{th} order feature vectors consisting of cepstral, delta and acceleration coefficients and log energy are extracted from 25 ms frames with 10 ms overlap. Whole word HMMs are trained with simple left-to-right models.

5. EXPERIMENTAL RESULTS

The original contributions of this work relate not to the optimisation but rather to an assessment of the fundamental limitations of spectral subtraction. The contribution of phase, cross-term and magnitude errors to the degradation in spectral subtraction performance is assessed as a function of SNR in terms of ASR word accuracy. Results are presented in Figure 1 for the Aurora 2 database in (a) and for the WSD(II) database in (b). The lowest profile in both figures corresponds to the baseline performance, i.e. without treatment by spectral subtraction. Without added noise baseline accuracies of 99% and 89% are observed for the two databases respectively. The differences in word accuracies at higher SNRs can be attributed to the different quantities of training data: 8440 for Aurora 2 c.f. 100 for WSD(II). At lower SNRs the superior performance of the WSD(II) database can be attributed to the hand-labelling of speech periods.

The first profiles illustrate the effect of phase errors. Using the corrupt speech phase to resynthesise the processed speech in the time domain, a negligible decrease in word accuracy is observed. At -5 dB phase errors cause a drop in word accuracy from 99% without added noise to 97% for the Aurora 2 database and from 89% without added noise to 86% for the WSD(II) database. Thus phase errors contribute very little to ASR performance degradation.

The second profiles illustrate the effect of cross-term errors which also appear to be negligible as the noise level increases to +5 dB. However, as the noise level increases further the degradation becomes non-negligible and, at -5 dB, the word accuracy falls to 90% and 66% for the Aurora 2 and WSD(II) databases respectively.

The next two profiles illustrate the performance of spectral subtraction with combined phase and cross-term errors, first with the actual noise values, $|N(e^{j\omega})|$ (third profile), and then with estimates, $|\hat{N}(e^{j\omega})|$ (fourth profile). Therefore the third profiles illustrate the fundamental limitations of spectral subtraction, given that phase and cross-term errors are generally ignored and indicate the likely optimal performance if a perfect estimate of the noise magnitude is applied. Word accuracies of 97% and 81% at +5 dB fall to 85% and 62% at -5 dB for the Aurora 2 and WSD(II) databases respectively. The profiles show that combined phase and cross-term errors again lead to negligible degradations for higher SNRs but that this increases to non-negligible levels as the SNR falls below 0 dB.

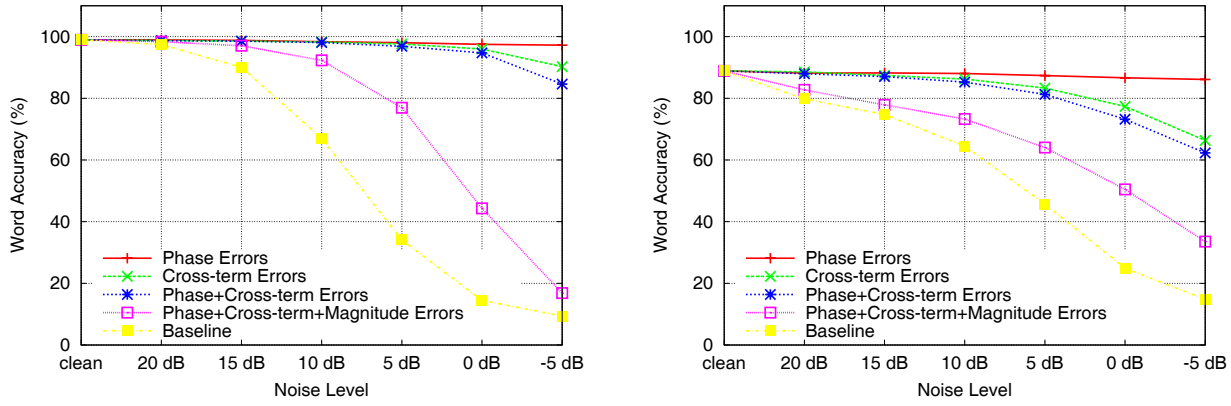
The objective now is to assess the performance of spectral subtraction in a practical scenario with a full complement of errors. The experiments thus relate to realistic conditions except perhaps that there is a constant, controlled SNR for each experiment. The fourth profiles confirm that the greatest contribution to ASR performance degradation is from magnitude errors. For the Aurora 2 database a word accuracy of 97% without magnitude errors falls to 77% at +5 dB and at -5 dB, a word accuracy of 85% without magnitude errors falls to 17% with magnitude errors. The differences are not so great for the WSD(II) database: 81% falls to 64% at +5 dB and 62% falls to 34% at -5 dB.

6. DISCUSSION

The two graphs in Figure 1 compare the contribution to ASR performance degradation when noisy speech is processed by spectral subtraction with phase, cross-term and magnitude errors. In both figures the top four profiles illustrate the degradation in ASR performance as each error is introduced. Magnitude errors are confirmed to produce greater degradations in ASR performance than phase and cross-term errors but, perhaps contrary to popular belief, as noise levels in the order of 0 dB are approached and exceeded, phase and cross-term errors can also make contributions that are not negligible.

The degradations in ASR performance caused by phase and cross-term errors is significantly less pronounced for the Aurora 2 database than for the WSD(II) database. This difference could be attributed to the greater quantity of training data for the Aurora 2 database which ensures a lesser degradation as phase and cross-term errors are introduced. However, given that noise estimates come from QBNE during both speech and non-speech intervals, the noise estimate is comparatively poor and as magnitude errors are introduced the degradation in ASR performance is more rapid. This would suggest that the contribution of magnitude errors is exaggerated in the case of results pertaining to the Aurora 2 database and that consequently the contribution of phase and cross-term errors might be proportionately greater than that illustrated.

The performance of speech enhancement in an ASR context is often gauged against the performance under clean conditions. For



(a) Aurora 2 Database

(b) WSD(II) Database

Fig. 1. ASR word accuracies for (a) the Aurora 2 database and (b) the WSD(II) database. Profiles illustrate, from top to bottom, ASR performance with phase errors (first profile), cross-term errors (second profile), combined phase and cross-term errors (third profile), combined phase, cross-term and magnitude errors (fourth profile) and for the baseline without treatment by spectral subtraction (fifth profile).

spectral subtraction, whilst this comparison is reasonable, it does not take into account the fundamental limitations that this experimental work highlights. In the application of spectral subtraction to speech enhancement considered here, unless the phase and cross-term errors are taken into consideration, ASR performance following spectral subtraction is likely to fall short of that under clean conditions, even with a perfect estimate of the noise magnitude.

7. CONCLUSIONS

There are three fundamental sources of error in a conventional implementation of spectral subtraction, namely phase, cross-term and magnitude errors. However, research efforts since the debut of spectral subtraction in 1979 often focus on obtaining the best possible estimates of the noise magnitude and neglect the remaining sources of error. This paper empirically highlights both the fundamental limitations of spectral subtraction and also the potential benefit of recently published works (e.g. [4–6]) that account for additional sources of error in the spectral subtraction model. Experimental results based on two standard European databases and ASR protocols confirm that errors in the magnitude do indeed make the greatest contribution to ASR performance degradation. However, as noise levels in the order of 0 dB are approached and exceeded the contributions of phase and cross-term errors are apparent and lead to non-negligible degradations in ASR performance. Future noise compensation and speech enhancement research should thus not only consider improved approaches to noise estimation in magnitude terms but also phase and cross-term errors, particularly at poor SNRs.

8. REFERENCES

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction," *IEEE Trans. ASSP*, vol. 27(2), pp. 113–120, 1979.

[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*, 1991, vol. 1, pp. 79–82.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," in *Proc. of the IEEE*, 1979, pp. 1586–1604.

[4] H. Xu, Z-H. Tan, P. Dalsgaard, and B. Lindberg, "Spectral Subtraction with Full-wave Rectification and Likelihood Controlled Instantaneous Noise Estimation for Robust Speech Recognition," in *Proc. Interspeech*, 2004, pp. 2085–2088.

[5] L. Deng, J. Droppo, and A. Acero, "Enhancement of Log Mel Power Spectra of Speech using a Phase-Sensitive Model of the Acoustic Environment and Sequential Estimation of the Corrupting Noise," *IEEE Trans. SAP*, vol. 12, pp. 133–143, 2004.

[6] Q. Zhu and A. Alwan, "The Effect of Additive Noise on Speech Amplitude Spectra: a Quantitative Analysis," *IEEE Signal Processing Letters*, vol. 9, pp. 275– 277, 2002.

[7] N. W. D. Evans, J. S. Mason, W. M. Liu, and B. Fauve, "On the Fundamental Limitations of Spectral Subtraction: An Assessment by Automatic Speech Recognition," in *Proc. EUSIPCO*, 2005.

[8] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," in *Proc. ICASSP*, 1979, pp. 208–211.

[9] D. Pearce and H. G. Hirsch, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in *Proc. ICSLP*, 2000, vol. 4, pp. 29–32.

[10] R. J. Jones, J. S. D. Mason, R. O. Jones, L. Helliker, and M. Pawlewski, "SpeechDat Cymru: A large-scale Welsh telephony database," in *Proc. LREC Workshop: Language Resources for European Minority Languages*, 1998.

[11] V. Stahl, A. Fischer, and R. Bippus, "Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering," in *Proc. ICASSP*, 2000, vol. 3, pp. 1875–1878.

[12] N. W. D. Evans and J. S. Mason, "Computationally Efficient Noise Compensation for Robust Automatic Speech Recognition Assessed under the AURORA 2/3 Framework," in *Proc. ICSLP*, 2002, vol. 1, pp. 485–488.