ON THE FUNDAMENTAL LIMITATIONS OF SPECTRAL SUBTRACTION: AN ASSESSMENT BY AUTOMATIC SPEECH RECOGNITION

Nicholas W. D. Evans, John S. 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}@swansea.ac.uk
web: http://eegalilee.swan.ac.uk

ABSTRACT

Spectral subtraction is one of the earliest and longest standing, popular approaches to noise compensation and speech enhancement. A literature search reveals an abundance of recent research papers that report the successful application of spectral subtraction to noise robust automatic speech recognition (ASR). However, as with many alternative approaches, the benefits lessen as noise levels in the order of 0 dB are approached and exceeded.

Previously published works relating to spectral subtraction provide a theoretical analysis of error sources. Recently the first empirical assessment showed that these fundamental limitations can lead to significant degardations in ASR performance. Results illustrate that under particularly high noise conditions these degradations are comparable to those caused by errors in the noise estimate which are widely believed to have by far the greatest influence on spectral subtraction performance. The original contribution made in this paper is the assessment of the fundamental limitations of a practiclal implmentation of spectral subtraction under the European standard ETSI Aurora 2 experimental protocols. Results illustrate that, perhaps contrary to popular belief, as noise levels in the order of 0 dB are approached phase and cross-term error sources do indeed contribute non-negligible degradations to ASR performance. This is believed to be a new observation in the context of spectral subtraction and ASR.

1. INTRODUCTION

The removal of background noise from speech signals has long been the subject of research interest and there exist a plethora of different approaches to accomplish what is often an extremely challenging task. Spectral subtraction is one of the earliest and longest standing approaches to noise compensation and speech enhancement brought about, in part, due to its simplicity and versatility. Spectral subtraction was developed in 1979 by Boll [1] and a literature search reveals an abundance of research papers, both long past and recent, that have investigated the application of spectral subtraction as well as the optimisation of the algorithm itself.

Depending on the application the assessment of spectral subtraction can be subjective (judged by human listeners) such as in the original work of Boll [1] and Berouti *et al* [2], or it can be objective in terms of automatic speech recognition (ASR) as in the subsequent work of Lockwood *et al* [3, 4, 5] from 1991. The majority of recent literature with a spectral subtraction theme focuses on ASR applications, assessing the effectiveness of the process in terms of ASR word accuracy with different spectral subtraction configurations.

However, improvements in ASR performance obtained through spectral subtraction 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.

Previously published works [6, 7, 8] relating to spectral subtraction provide a theoretical analysis of error sources, namely phase, cross-term and magnitude errors. Surprisingly perhaps, to the Authors' best knowledge there are no studies that have compared the contribution of each error source to the performance of spectral subtraction using controlled, standard experimental conditions. Thus herein lie the objectives of this paper, namely to assess each individual error source as a function of noise level. Experiments are performed in a conventional spectral subtraction framework where the cost function is ASR word accuracy.

The remainder of this paper is organised as follows. Section 2 describes what might be considered as a conventional implementation of spectral subtraction and also describes the error sources and fundamental limitations of spectral subtraction that are assessed in this paper. Section 3 describes the ASR database and experimental setup. Results are presented in Section 4 and conclusions follow in Section 5.

2. SPECTRAL SUBTRACTION

Spectral subtraction is not a recent approach to noise compensation and was first proposed in 1979 [1]. There is however a vast amount of more recent work in the literature relating to different implementations and configurations of spectral subtraction. The objective in this section is thus to describe what is perhaps best termed as a conventional implementation of spectral subtraction drawing from [1, 2, 3] and is that upon which the experimental work in this paper is based. Section 2.1 describes the implementation and Section 2.2 illustrates the fundamental limitations of the approach.

2.1 Implementation

The goal of spectral subtraction is the suppression of additive noise from a corrupt signal, in this case a speech 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 *c*lean speech (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. The important point is that phase terms are ignored. Both forms of *magnitude* subtraction occur frequently in the literature and perhaps for practical reasons little or no reference is made to phase. Power (magnitude) subtraction is adopted here as it is more common in the literature and since experimental evidence suggests there is little difference between the two [9].

The noise estimate in Equation 2 is conventionally obtained during non-speech intervals and in the frequency domain from short term magnitude spectra:

$$|\hat{N}(e^{j\omega})|^2 = \frac{1}{T}\sum_{i=1}^{T}|D_i(e^{j\omega})|^2, \quad (3)$$

where, $|D(e^{j\omega})|^2$ is the observed signal and where, for example in [1], $i = 1...T$ corresponds to an average over 1/3 s.

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, $\theta_D(e^{j\omega})$:

$$\hat{S}(e^{j\omega}) = \Big[\ |D(e^{j\omega})|^2 - |\hat{N}(e^{j\omega})|^2\ \Big]^{1/2} e^{\theta_D(e^{j\omega})} \quad (4)$$

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 has found that improved results are obtained by employing noise over-estimates and noise floors, the ideas for which were introduced by the early original work of Berouti [2]. Equation 4 is thus modified to:

$$\begin{aligned}\hat{S}(e^{j\omega}) \quad = \quad & \max\Big(\Big[\ |D(e^{j\omega})|^2 - \alpha|\hat{N}(e^{j\omega})|^2\ \Big], \\ & \beta|D(e^{j\omega})|^2\Big)^{1/2} e^{\theta_D(e^{j\omega})}, \end{aligned} \quad (5)$$

where α is the noise over-estimation parameter and β is the noise floor as in [2, 3]. The idea is to artificially increase noise attenuation through α and then to simultaneously suppress musical noise and negative values in the processed magnitude spectrum through β. The two parameters are usually noise-dependent, an intuitive illustration of which is provided by considering spectral subtraction as a zero-phase filter and plotting the gain against the noisy-signal-to-noise-ratio (NSNR) as in [8, 9].

2.2 Fundamental Limitations

The emphasis here is to illustrate the fundamental limitations of spectral subtraction. In [6] it is shown that the clean speech spectrum, $S(e^{j\omega})$, in exact terms, is expressed by:

$$\begin{aligned}S(e^{j\omega}) \quad = \quad & \Big[\ |D(e^{j\omega})|^2 - |N(e^{j\omega})|^2 - S(e^{j\omega})\cdot N^*(e^{j\omega}) - \\ & S^*(e^{j\omega})\cdot N(e^{j\omega})\ \Big]^{1/2} e^{\theta_S(e^{j\omega})}, \end{aligned} \quad (6)$$

where $S(e^{j\omega})\cdot N^*(e^{j\omega})$ and $S^*(e^{j\omega})\cdot N(e^{j\omega})$ are termed throughout this paper as cross-terms. Comparing Equations 4 and 6 there are thus three sources of error in a practical implementation of spectral subtraction:

- phase errors, errors arising from the differences between $\theta_S(e^{j\omega})$ and $\theta_D(e^{j\omega})$,
- cross-term errors, from neglecting $S(e^{j\omega})\cdot N^*(e^{j\omega})$ and $S^*(e^{j\omega})\cdot N(e^{j\omega})$, and
- magnitude errors, which refer to the differences between $|N(e^{j\omega})|$ and $|\hat{N}(e^{j\omega})|$.

It is usually assumed that phase errors do not impact on ASR performance. Clearly returning to a time domain representation of the processed speech is likely to introduce phase errors, associated with the differences between $\theta_S(e^{j\omega})$ and $\theta_D(e^{j\omega})$. Phase errors are considered in this paper to embrace situations where it is desirable to produce an enhanced time domain speech signal as well as ASR. Cross-terms are also thought to have a negligible effect based on the assumption that the speech and noise are uncorrelated, thus in discussion and analysis they are generally omitted. The procedure in practice focuses only on the magnitude: obtaining effective estimates of $|N(e^{j\omega})|$.

It is the objective of the experimental work presented in this paper to assess the impact on ASR of these fundamental assumptions. First though, the experimental database and ASR configuration is described.

3. DATABASE AND ASR CONFIGURATION

The European standard Welsh SpeechDat(II) FDB-2000 database [10], hereafter referred to as WSD(II), is used throughout with a standard ETSI Aurora 2 style experimental setup [11, 12].

3.1 The Welsh SpeechDat(II) Database

The WSD(II) telephony database was collected largely over a public switched telephone network and a smaller component over various cellular networks in the UK. The motivations for using this database, apart from its obvious suitability as a telephony ASR database, arose through the preference for a labelled database. The labelling means that noise estimation in non-speech intervals is easily implemented without the prior optimisation of a voice activity detector (VAD). Assessment of spectral subtraction through ASR is not then influenced by the performance of a VAD or alternative approaches to noise estimation.

3.2 Speech Data and Noise Addition

The WSD(II) database was recorded from members of the Welsh speaking public. The database contains typical significant levels of home background noise with a smaller number of more noisy mobile phone calls. The interest here is in assessing the limitations of spectral subtraction in the presence of noise and so further noise was added to the original clean speech. For the experimental results reported in this paper car noise was added to the clean speech data justified by the popular application of in-car, noise robust automatic speech recognition.

From the mobile telephony components of the database, a subset of 10 Welsh isolated digits was selected comprising 100 speaker training utterances and 1500 speaker test utterances; each speaker contributes only one utterance (either test or training). There is no overlap between speakers in training and testing. To evaluate the performance of spectral subtraction under different levels of noise, real car noise was added to the test data at six different SNRs (20, 15, 10, 5, 0 and -5 dB), as is the case with the Aurora 2 database. Noise addition was performed using standard ITU software conforming to the G.712 [13] and P.56 [14] standards and again follows closely the experimental setup of the Aurora 2 database. One difference in the setup is that the speech data in the WSD(II) database was collected 'in the field' and not under laboratory 'clean' conditions, as is the case for the original TIDigits data [15] of the Aurora 2 database. Consequently, the G.712 filtering that is applied to the speech and noise signals in the Aurora 2 setup, was applied only to the noise in the WSD(II) setup, the speech data having already been telephony-band filtered. No noise is added to the training data.

3.3 Feature Extraction and Recognition

The Aurora 2 W1007 standard front-end [11, 12] is used for feature extraction and an Aurora 2 style back-end, modified to utilise speech, non-speech labels is used for recognition. Details of the Aurora 2 front-end and the HTK reference recogniser can be found in [11, 12]. 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. As for the Aurora 2 standard experimental setup, whole word HMMs are trained with simple left-to-right models.

The baseline WSD(II) recognition results are presented in Figure 1 (last profile). A word recognition accuracy of 89% without added noise drops to 15% at -5 dB. Spectral subtraction is adopted

as a pre-processing, speech enhancement stage prior to feature extraction. All improvements may therefore be attributed to spectral subtraction and not to any modifications to either the feature extraction or recognition stages.

4. EXPERIMENTAL WORK

The objectives and original contribution of this work relate not to the optimisation of spectral subtraction but rather to an assessment of the fundamental limitations of spectral subtraction. The contribution of each error source to the degradation in spectral subtraction performance are compared first independently and then collectively in terms of ASR word accuracy. The experimental work was performed under the standard experimental conditions outlined above and in a common, conventional spectral subtraction implementation described above and in Section 4.1. Phase errors are assessed in Section 4.2, cross-term errors in 4.3 and magnitude errors in 4.4.

4.1 Spectral Subtraction Framework

Each error source is assessed with a common spectral subtraction implementation. The complex, frequency domain representations of the clean speech, $S(e^{j\omega})$, corrupt speech, $D(e^{j\omega})$, and corresponding noise, $N(e^{j\omega})$, are 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 6 are all known and thus the contribution to ASR performance degradation due to phase and cross-term errors may be assessed independently and collectively. Noise estimation is performed over 0.5 s during non-speech intervals either side of speech periods. In each condition the noise over-estimate, α, and noise floor, β, are varied as indicated below and chosen to optimise ASR word accuracy for each noise level.

4.2 Phase Errors

Modified to utilise noise over-estimates and noise floors and to include phase errors, Equation 6 is rewritten as:

$$\hat{S}(e^{j\omega}) = \max\Big(\Big[\,|D(e^{j\omega})|^2 - \alpha|N(e^{j\omega})|^2 - S(e^{j\omega})\cdot N^*(e^{j\omega}) - S^*(e^{j\omega})\cdot N(e^{j\omega})\,\Big], \beta|D(e^{j\omega})|^2\Big)^{1/2} e^{\theta_D(e^{j\omega})} \quad (7)$$

The first profile in Figure 1 illustrates the effect of phase errors on ASR performance as a function of SNR. Using the corrupt speech phase to resynthesise the processed speech in the time domain, a negligible decrease in word error rate is observed. Phase errors cause a drop in word accuracy from 89% under clean conditions to 86% at -5 dB. Thus phase errors contribute very little to ASR performance degradation.

4.3 Cross-term Errors

Cross-terms are also commonly assumed to have only a small influence on spectral subtraction performance. This is because in the ideal the cross-term components, $S(e^{j\omega})\cdot N^*(e^{j\omega})$ and $S^*(e^{j\omega})\cdot N(e^{j\omega})$, average to zero.

The known noise values are again used for subtraction but cross-term components are omitted. The processed signal is then resynthesised with the phase of the original speech, $\theta_S(e^{j\omega})$:

$$\hat{S}(e^{j\omega}) = \max\Big(\Big[\,|D(e^{j\omega})|^2 - \alpha|N(e^{j\omega})|^2\,\Big], \beta|D(e^{j\omega})|^2\Big)^{1/2} e^{\theta_S(e^{j\omega})} \quad (8)$$

The second profile in Figure 1 illustrates ASR performance with cross-term errors as a function of SNR. The performance degradation caused by cross-term errors becomes significant at the lowest SNRs. At 10 dB a word accuracy of 86% is observed. This falls to 66% at -5 dB.

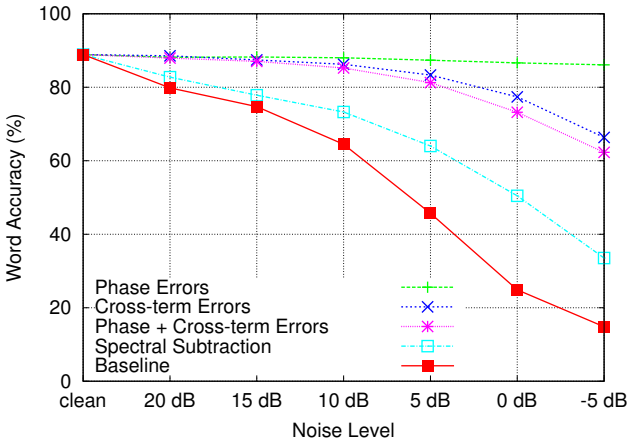

Figure 1: ASR word accuracy for the WSD(II) database with different sources of error in a common spectral subtraction implementation. The five profiles illustrate, from top to bottom, ASR performance with phase errors (first profile), cross-term errors (second profile), combined phase and cross-term errors (third profile), and the performance of conventional spectral subtraction with all three errors sources (fourth profile). Profiles illustrated together with the baseline performance without treatment by spectral subtraction (fifth profile).

4.4 Magnitude Errors

Two further profiles in Figure 1 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).

The third profile illustrated in Figure 1 illustrates the fundamental limitations of spectral subtraction, given that by convention, phase and cross-term errors are ignored. The profile therefore illustrates the likely optimal performance if, in a conventional implementation of spectral subtraction, a perfect estimate of the noise magnitude, $|N(e^{j\omega})|$, is applied. The subtraction is thus as in Equation 8 except that $e^{\theta_S(e^{j\omega})}$ is replaced by $e^{\theta_D(e^{j\omega})}$. At 10 dB a word accuracy of 85% is observed. This falls to 62% at -5 dB. The profiles show that phase and cross-term errors lead to relatively negligible degradations in ASR performance for higher SNRs but that this increases to non-negligible levels as SNRs in the order of 0 dB are approached.

A configuration with conventional noise estimates is now considered. The subtraction now incorporates a full complement of errors: phase, cross-term and magnitude errors as per Equation 4. The experiments thus relate to realistic conditions except perhaps that there is a constant, controlled SNR for each experiment. The objective is to compare performance with a full complement of errors to conventional spectral subtraction with a perfect noise estimate. The profiles show that for SNRs above 0 dB the greatest contribution to ASR performance degradation comes from magnitude errors and, at 0 dB, corresponds to a word accuracy of 75% without magnitude errors to 51% with magnitude errors. However, as SNRs of 0 dB are exceeded the contribution of phase and cross-term errors increases to a comparable level to that of magnitude errors. At -5 dB a word accuracy of 62% without magnitude errors falls to 35% with a full complement of errors.

4.5 Discussion

Figure 1 compares the contribution to ASR performance degradation coming from phase (Section 4.2), cross-terms (Section 4.3) and combined errors (phase, cross-term and magnitude errors, Section 4.4). The top four profiles illustrate the degradation in ASR performance as each error is introduced. The top two profiles illus-

trate ASR performance with only phase errors and only cross-term errors respectively. The third profile illustrates performance with combined phase and cross-term errors and thus represents a conventional implementation of spectral subtraction though with a perfect noise estimate. The fourth profile (spectral subtraction) illustrates performance with combined phase, cross-term and magnitude errors and thus represents the performance of spectral subtraction in a realistic sense. Except for the lowest noise levels, magnitude errors are confirmed to lead to significantly greater degradations in ASR performance than phase and cross-term errors. These results illustrate that, perhaps contrary to popular belief, as noise levels in the order of 0 dB are approached and exceeded phase and cross-term errors do indeed contribute to ASR performance degradation on a scale comparable to the degradations caused by errors in the magnitude. This is believed to be a new observation in the context of spectral subtraction and ASR.

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

5. CONCLUSIONS

Research efforts since the debut of spectral subtraction in 1979 often focus on obtaining the best possible estimates of the noise magnitude. Previously published work has identified three error sources in a conventional implementation of spectral subtraction, namely phase, cross-term and magnitude errors. This is believed to be the first paper to assess the fundamental limitations of spectral subtraction in a conventional implementation through ASR and controlled, standard experimental conditions. Results confirm that, except for the worst levels of SNR, 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 the contributions of phase and cross-term errors are not negligible and lead to degradations that are comparable to those caused by magnitude errors. This observation indicates that new approaches to noise compensation and speech enhancement should perhaps consider phase and cross-term errors, particularly at poor SNRs.

REFERENCES

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

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

[3] 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.

[4] 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*, 1991, vol. 1, pp. 83–86.

[5] P. Lockwood and J. Boudy, "Experiments with a Nonlinear Spectral Subtraction (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars," *Speech Communiaction*, vol. 11, pp. 215–228, 1992.

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

[7] S. M. McOlash, R. J. Niederjohn, and J. A. Heinen, "A Spectral Subtraction Method for the Enhancement of Speech Corrupted by Non-white, Non-stationary Noise," *Proc IEEE. Int. Conf. on Industrial Electronics, Control, and Instrumentation*, vol. 2, pp. 872–877, 1995.

[8] X. Huang, A. Acero, and H-W. Hon, *Spoken Lanugage Processing*, Prentice Hall, 2001.

[9] N. W. D. Evans, "Spectral Subtraction for Speech Enhancement and Automatic Speech Recognition," *PhD Thesis, University of Wales Swansea*, 2003.

[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] H. G. Hirsch and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions," *ISCA ITRW ASR2000 'Automatic Speech Recognition: Challenges for the next Millenium'*, 2000.

[12] 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.

[13] ITU recommendation G.712, *Transmission performance characteristics of pulse code modulation channels*, ITU, 1996.

[14] ITU recommendation P.56, *Objective measurement of active speech level*, ITU, 1993.

[15] R. G. Leonard, "A database for speaker independent digit recognition," in *Proc. ICASSP*, 1984, vol. 3, pp. 42.11–14.