A Modified LIMA Framework for Spectral Subtraction Applied to In-Car Speech Recognition

Tristan Kleinschmidt, Sridha Sridharan, Michael Mason
Speech and Audio Research Laboratory
Queensland University of Technology,
GPO Box 2434, Brisbane, Australia, 4001
{t.kleinschmidt, s.sridharan, m.mason}@qut.edu.au

***Abstract*— In noisy environments, speech recognition accuracy degrades significantly. Speech enhancement algorithms have been designed to overcome this, however solutions to date have not been optimal for speech recognition especially for non-stationary noise like that in a car. Recently, a likelihood-maximising (LIMA) criteria has been applied to speech enhancement techniques. This paper analyses the suitability of spectral subtraction for potential use under a modified version of this framework where direct access to and manipulation of speech recognition models is not available. Analysis shows spectral subtraction is suited to this holistic LIMA approach by confirming the cost surface is appropriate for gradient descent methods. It is also observed that there are regions on the cost surface where performance exceeds that achieved by parameter values traditionally selected for spectral subtraction.**

I. Introduction

A key challenge of deploying speech recognition in real-world environments is the requirement to perform well in the presence of high levels of noise. Since most speech recognition systems are trained for use in controlled environments, they fail to produce satisfactory performance under more adverse conditions.

Methods for robust speech recognition include model compensation, use of robust features and recognition algorithms, as well as speech enhancement. Enhancement is a popular approach as little-or-no prior knowledge of the operating environment is required for improvements in recognition accuracy.

Popular speech enhancement algorithms (e.g. filter-and-sum beamforming or spectral subtraction) have been primarily designed to improve intelligibility and/or quality of the speech signal without consideration of what effect that may have on other speech processing systems [1]. Optimisation in these algorithms is focussed on signal-based measures including maximising signal-to-noise ratio or minimisation of the mean-squared signal error. Some of these techniques produce improvements in word accuracy performance, but these improvements are by-products, rather than the goals of the enhancement techniques.

One possible solution to the problem is to use speech recognition likelihoods as the optimisation criteria in the enhancement algorithms. Promising results have been shown in recent studies using this approach [1], [2], [3], [4]. In their current form these techniques require access to the underlying state models and attempt to jointly optimise both state sequences and enhancement parameters. This paper proposes a modified approach in which the speech recogniser can be regarded as a ‘black-box’. This approach removes the need for access to the recogniser’s acoustic models and a fully decoded state sequence. The details of this approach and its applicability to use with spectral subtraction is presented. Spectral subtraction is chosen for its simplicity and common use in single-channel speech enhancement applications.

The rest of this paper is presented as follows. Section II provides background on spectral subtraction speech enhancement. Section III looks at the likelihood-maximising (LIMA) framework and its application to spectral subtraction. Preliminary experimental results and discussion of the importance of these results is presented in Section IV.

II. Spectral Subtraction

In a noisy environment, speech $s(n)$ is assumed to be corrupted by additive background noise $d(n)$ to produce corrupted speech $y(n)$ as follows:

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

Equation (1) can be represented in frequency domain as:

$$Y(\omega) = S(\omega) + D(\omega) \tag{2}$$

Generally, an estimate of the magnitude (or power) spectra of the noise signal $\hat{D}(\omega)$ is subtracted from the corresponding spectra of the noisy signal $Y(\omega)$ to give an estimate of the clean speech signal $\hat{S}(\omega)$:

$$|\hat{S}(\omega)|^\gamma = |Y(\omega)|^\gamma - |\hat{D}(\omega)|^\gamma \tag{3}$$

where γ is the power exponent which equals 1 for magnitude spectral subtraction or 2 for power spectral subtraction [5]. The phase component of the noisy speech signal is left unaltered and is kept for reconstruction into the time domain.

Should the subtraction in (3) give negative values (i.e. the noise estimate $|\hat{D}(\omega)|^\gamma$ is greater than the signal $|Y(\omega)|^\gamma$)

Parts of the work presented here were funded through the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

a flooring factor is introduced. This leads to the following formulation of spectral subtraction:

$$|\hat{S}(k)|^\gamma = \begin{cases} |Y(k)|^\gamma - |\hat{D}(k)|^\gamma & |\hat{D}(k)|^\gamma > |Y(k)|^\gamma \\ \beta|\hat{D}(k)|^\gamma & \text{otherwise} \end{cases} \quad (4)$$

where β is the noise floor factor, and $0 < \beta \ll 1$ [5]. Common values for this parameter range between 0.005 and 0.1 [5], [6].

Although common values for γ and β are those noted above, there is actually no limitation on the values that these parameters can take. These values are typically used for their conceptual meanings as opposed to performance. Altering these two parameters can make considerable difference to the speech recognition performance of spectral subtraction, as will be demonstrated in Section 4.

It should also be noted here that in order to derive the two common rules denoted in (3) two conflicting assumptions are made. If the clean speech and noise signals are assumed to be uncorrelated, the power spectral subtraction rule (i.e. $\gamma = 2$) results. Alternatively, if the two signals are assumed to be co-linear, the equation reduces to the magnitude spectral subtraction rule. In practice, neither of these assumptions is valid all the time. This leads to the possibility of optimising these parameters to best fit the instantaneous relationship between clean speech and noise signals.

III. LIMA Spectral Subtraction

As mentioned in Section I, in recent studies the likelihood-based criterion has been used to replace traditional signal-level criteria in speech enhancement algorithms with the aim to improve speech recognition accuracies. This was seen to minimise distortion in the *effective* auditory signal for recognition purposes instead of the distortion of the speech waveform [3]. Techniques which maximise the likelihood in the speech recogniser are referred to as LIMA enhancement techniques.

The LIMA framework first generates an initial state sequence using the speech recogniser. This sequence is used to optimise the parameters using a gradient-descent algorithm – ensuring an optimal set of parameters for the proposed state sequence. The utterance is decoded again using the new parameters to generate a new state sequence. This joint optimisation of both the array parameters and state sequence continues until the recognition likelihood converges.

Formulated in this manner, it is required to obtain both frame-by-frame state sequences and access to the model set in order to perform optimisation in LIMA techniques. This paper proposes a modification to the LIMA framework aimed at removing the need for access to state models *and* state sequence information – information rarely available when endeavouring to integrate third party recognition engines in practical applications. Here, we assume that only access to full utterance likelihoods and word sequences is available.

Using these two pieces of information, a ”blind” gradient-descent approach can be applied whereby a new set of enhancement parameters are tried and the resulting likelihood compared to that of the previous iteration. The comparison directs the correct direction to take.

This method may be seen to be more restrictive than the original framework as it may require a series of enhancement and recognition steps in order to determine a valid direction of optimisation. This is not the case in the existing work as the optimisation takes place directly on the state sequence.

The modified framework does however ensure no internal information about the recogniser is required. It may also remove some of the reliance on the initial state sequence which is a downfall of the existing framework.

In order to apply the modified (or original) framework to spectral subtraction the two parameters referred to in Section II constitute the full parameter set. We denote this set by:

$$\xi_{ss} = [\gamma, \beta] \quad (5)$$

An investigation into the affect of altering this parameter set is presented in Section IV.

IV. Experimental Results

To evaluate the suitability of the modified LIMA framework, two experiments were designed using spectral subtraction as the enhancement method of interest. The first experiment investigated the existence of enhancement parameters which provided superior performance to traditionally selected values of γ and β in spectral subtraction. The second experiment extends the initial experiment by examining whether a gradient-descent method would still be appropriate for optimising the enhancement parameter set when only full utterance scores were available.

Both experiments use speaker-independent, context-dependent 3-state triphone Hidden Markov Models (HMM) trained using the Wall Street Journal 1 corpus. The models were trained using 39-dimensional Mel-Frequency Cepstral Coefficient (MFCC) vectors - 13 MFCC (including C_0) plus delta and acceleration coefficients. Each HMM state was represented using a 16-component Gaussian Mixture Model.

Experimental data came from the phone numbers task of the AVICAR database collected by the University of Illinois [7]. This database contains real speech recordings under 5 different driving conditions: idle (IDL), 35mph with windows up (35U) and down (35D), and 55mph with windows up (55U) and down (55D). In this way, performance under specific noise conditions is of interest as opposed to different signal-to-noise ratios. Microphone number 4 of the 8-channel recordings was utilised. The first experiment included utterances from 61 distinct speakers (30 male, 31 females) and the second used utterances from 20 speakers (14 male, 6 female).

A. Experiment 1

In order to demonstrate that varying the values of the two parameters in (5) alters speech recognition accuracy, a number of recognition experiments were conducted. A selection of 3140 phone number utterances from the test database were used.

TABLE I

WORD RECOGNITION ACCURACIES (%) FOR VARYING VALUES OF SPECTRAL SUBTRACTION PARAMETERS.

	IDL	35U	35D	55U	55D
Baseline	75.24	49.95	36.18	41.00	22.35
γ=1.0,β=0.1	80.70	47.34	37.42	39.62	27.82
γ=1.5,β=0.1	81.37	51.77	41.16	44.65	29.12
γ=2.0,β=0.1	81.15	53.92	42.01	46.24	28.66
γ=1.5,β=0.1	81.37	51.77	41.16	44.65	29.12
γ=1.5,β=0.3	81.94	57.21	43.84	50.11	29.63
γ=1.5,β=0.5	80.37	56.86	42.62	48.93	27.59

Values for γ and β were varied in linear increments through the ranges [1.0, 2.0] and [0.1, 0.5] respectively. Word recognition accuracies for increments of γ by 0.5 and β by 0.2 and are shown in Table I.

It can be seen from the table that altering the spectral subtraction parameter values leads to changes in speech recognition performance and that it is possible to locate values of β and γ which provide better word recognition performance than those commonly proposed in literature - the accuracy at $\beta = 0.3$ and $\gamma = 1.5$ exceeds that achieved when $\beta \leq 0.1$ and $\gamma = 1 \,\text{or}\, 2$. These findings show the potential for spectral subtraction parameter optimisation under a LIMA framework.

B. Experiment 2

Evaluation of the potential for the modified LIMA framework to find optimal spectral subtraction parameters using gradient-descent methods was performed using a selection of 250 of the phone number utterances used in experiment 1. The values for γ and β were varied in linear increments through the ranges [1.0, 5.0] and [0.1, 3.0] respectively.

Fig. 1 shows a typical surface of recognition likelihood scores versus variations in γ and β. The general shape of the surface was observed to be common to all utterances tested, suggesting that it is utterance, speaker and noise-independent. We observe that increases in either γ or β within the ranges specified above leads to an increase in the likelihood score of the utterance. It can also be seen that whilst the likelihood surface flattens out considerably, it is still marginally increasing. From this figure, it is believed that the likelihood surface may be monotonically increasing, which is very problematic for gradient-descent optimisation.

To avoid this problematic feature of the cost surface, it is important to identify which likelihood scores are associated with correct transcriptions. Region 2 in Fig. 1 depicts the typical location and shape of the region associated with correct transcriptions. This region was observed to vary in size depending on speaker, utterance and noise level. Examples of the variations associated with noise level are depicted in Fig. 2. As the noise level increases (a-d) the size of the correct surface diminishes considerably, and changes shape slightly. This is expected as the increased levels of noise hamper the speech recogniser.

The results presented indicate that in order for the proposed modified LIMA framework to perform optimisation, it is important that the region of correct transcriptions is able to be identified. Whilst this may appear to be hidden information, we are aware of several voice control applications where utterance confirmation is a well established mechanism. Therefore by collecting user confirmations and associating them with the utterances of interest, there is sufficient information to perform the likelihood maximisation for the benefit of future utterances.

V. CONCLUSION

From the results of the second experiment, we believe that the modified LIMA framework, which attempts to optimise enhancement parameters based on whole utterance scores and without access to state sequences or models, is capable of being applied to a system using spectral subtraction. This conclusion is supported through the observation of a smooth cost surface suitable for gradient based optimisation.

The first experiment demonstrated that in addition to the proposed framework being able to blindly optimise spectral subtraction parameters using only utterance level scores, that there was also the potential to achieve better performance when the values of β and γ are not constrained to their traditionally used values.

REFERENCES

[1] M. Seltzer, B. Raj, and R. Stern, “Likelihood-maximizing beamforming for robust hands-free speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, 2004.

[2] A. Sankar and C.-H. Lee, “A maximum-likelihood approach to stochastic matching for robust speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, 1996.

[3] M. Seltzer and R. Stern, “Subband likelihood-maximizing beamforming for speech recognition in reverberant environments,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2109–2121, 2006.

[4] G. Shi, P. Aarabi, and H. Jiang, “Phase-based dual-microphone speech enhancement using a prior speech model,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 109–118, 2007.

[5] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 1979, pp. 208–211.

[6] R. Martin, “Spectral subtraction based on minimum statistics,” in *EUSIPCO*, Edinburgh, 1994, pp. 1182–1185.

[7] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, “Avicar: Audio-visual speech corpus in a car environment,” in *INTERSPEECH*, Jeju Island, Korea, 2004, pp. 2489–2492.

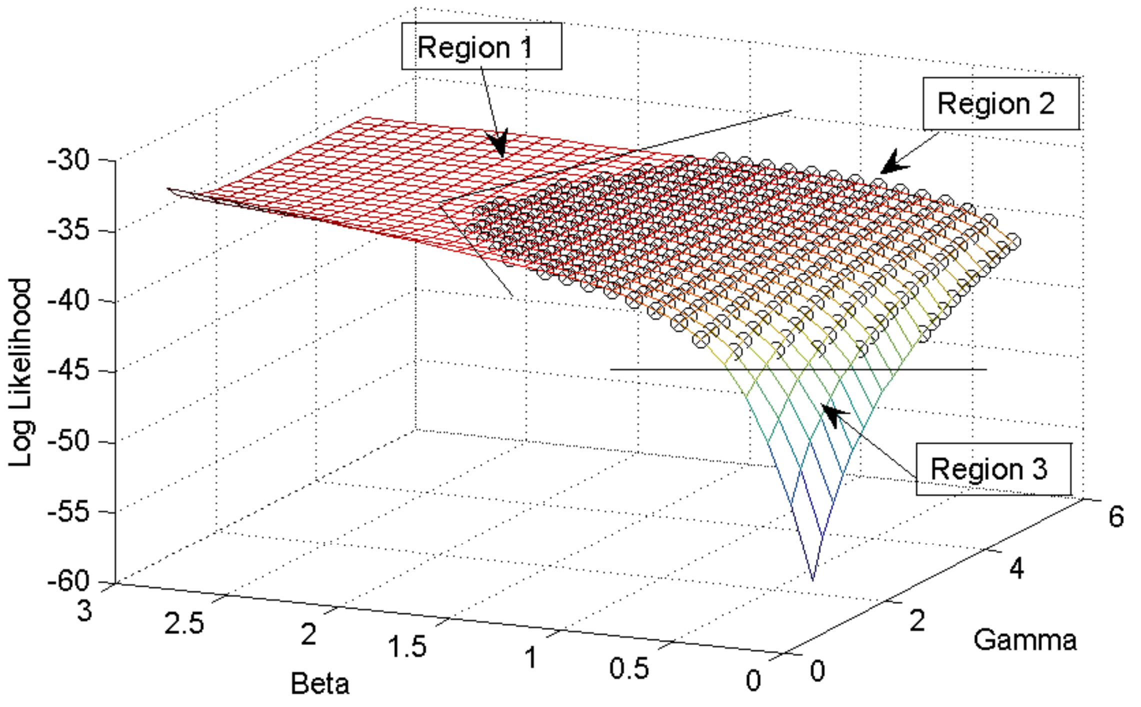
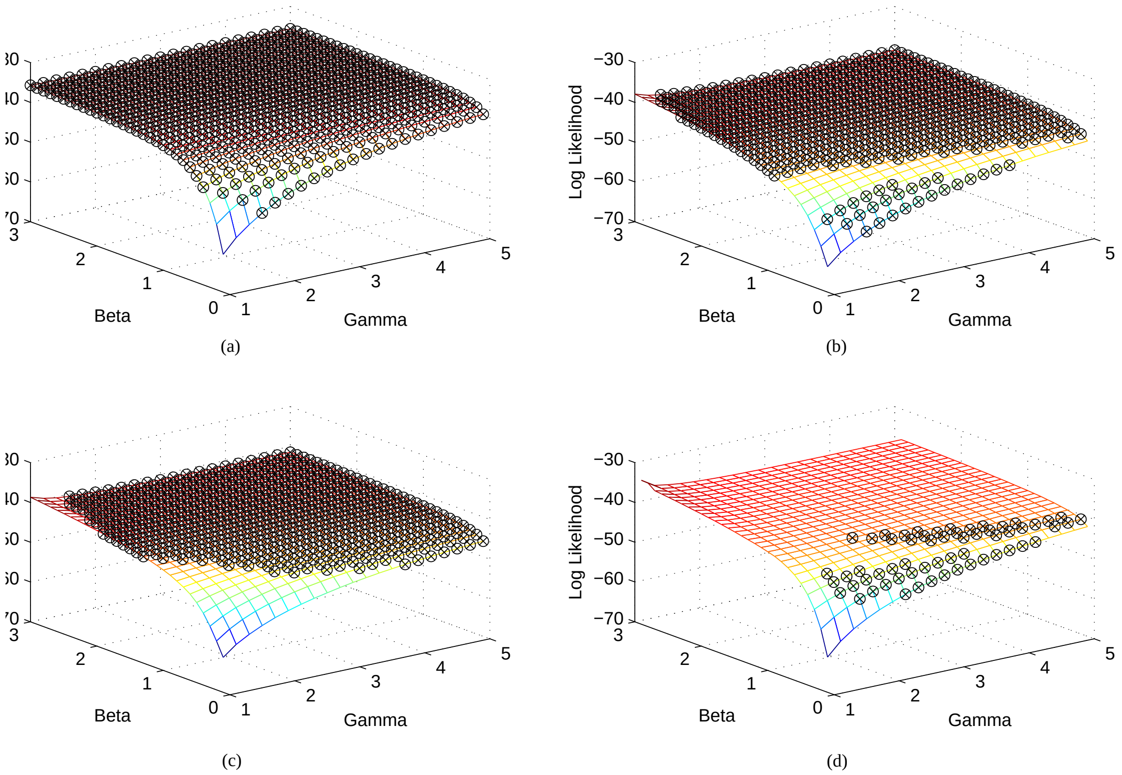


Fig. 1. *Visualisation of the likelihood surface for varying β and γ. Region 1 is a region of high distortion; Region 2 is the region of 100% accuracy; Region 3 exhibits insufficient speech enhancement to recover in speech recognition.*



(a)

(b)

(c)

(d)

Fig. 2. *An example of correct utterance region decreasing as noise levels increase. Noise levels in the figures are (a) -40.8dB, (b) -35.0dB, (c) -33.1dB, and (d) -23.6dB.*