SNR Features for Automatic Speech Recognition

Philip N. Garner
Idiap Research Institute
Martigny, Switzerland
pgarner@idiap.ch

***Abstract*—When combined with cepstral normalisation techniques, the features normally used in Automatic Speech Recognition are based on Signal to Noise Ratio (SNR). We show that calculating SNR from the outset, rather than relying on cepstral normalisation to produce it, gives features with a number of practical and mathematical advantages over power-spectral based ones. In a detailed analysis, we derive Maximum Likelihood and Maximum a-Posteriori estimates for SNR based features, and show that they can outperform more conventional ones, especially when subsequently combined with cepstral variance normalisation. We further show anecdotal evidence that SNR based features lend themselves well to noise estimates based on low-energy envelope tracking.**

I. Introduction

An important problem encountered in speech signal processing is that of how to normalise a signal for the effects of noise. In speech enhancement the task is to remove noise from a signal to reproduce the uncorrupted signal such that it is perceived by a listener to be less noisy. In Automatic Speech Recognition (ASR), the task is to reduce the effect of noise on recognition accuracy. In this paper, we will concentrate on the latter (ASR) problem.

Two categories of noise are generally considered: Additive noise is that which represents a distinct signal other than the one of interest. Convolutional noise is that which alters the spectral shape, and can be associated with either the signal of interest, or both the signal and the additive noise.

Cepstral Mean Normalisation (CMN) is a well established technique that compensates for convolutional noise. It is based on the persuasive observation that a linear channel distortion becomes a constant offset in the cepstral domain. CMN also affords some robustness to additive noise. Cepstral Variance Normalisation (CVN) has been observed to provide further noise robustness [1], and the combination of CMN and CVN is now quite ubiquitous in ASR.

Orthogonal to the cepstral normalisation approach, many common practical solutions for additive noise compensation are based on the assumption of a simple additive Gaussian model for both speech and noise in the spectral domain. In ASR, the spectral subtraction approach of Boll [2] is well established, and often used as a means to derive a Wiener filter. In speech enhancement, much work is based on the technique of Ephraim and Malah [3]. Both these techniques have influenced the design of the ETSI standard ASR front-end [4].

Techniques that rely on noise subtraction are dependent upon some means of measuring the background noise in a signal. Often, it is sufficient to simply average the first few frames of an utterance, however this is not robust to changing noise levels. Ris and Dupont [5] present a survey of methods to measure noise, favouring the low-energy envelope tracking approach of Martin [6]. Lathoud *et al.* [7] present a statistical spectral model that yields both noise and speech estimates.

Cepstral and spectral techniques are often combined. This is a natural approach as, theoretically, the two approaches are designed to tackle different types of noise. For instance, histogram normalisation, a logical progression of CMN/CVN to higher order moments, has been successfully combined with spectral compensation techniques by Segura *et al.* [8]. Lathoud *et al* [7], who describe their technique as "Unsupervised" spectral subtraction (USS), also report good results in combination with cepstral normalisation.

In this paper, we analyse the relationship between spectral and cepstral normalisation. We first present a simplistic analysis, then a more detailed Bayesian analysis, showing that knowledge of the presence of cepstral compensation should influence the chosen approach to spectral compensation. Theoretical results are evaluated leading to a conclusion that SNR based features represent a theoretically rigorous but computationally simple approach to ASR, and could easily be incorporated into more advanced techniques.

II. Simplistic approach to noise

A. Cepstral Mean Normalisation

In a simplistic, but informative, view of an ASR front-end, an acoustic signal is Fourier transformed to give a vector of spectral coefficients $(s_1, s_2, \ldots, s_F)^\mathsf{T}$. After a linear transform implementing a non-linear frequency warp, the cepstrum is calculated. The cepstrum involves a logarithm followed by another linear transform. In the presence of only convolutional noise, $(c_1, c_2, \ldots, c_F)^\mathsf{T}$, which is multiplicative in the frequency domain, the logarithm becomes

$$\log(c_f s_f) = \log(c_f) + \log(s_f), \tag{1}$$

where $\log(c_f)$ is constant over time, but $\log(s_f)$ varies. Hence, subtraction of the cepstral mean results in removal of the constant convolutional noise term. When the filter-bank is considered, the above holds if the c_f are assumed constant within a given filter-bank bin.

In the presence of only additive noise, the noise is assumed to remain additive after the Fourier transform. In this sense,

the logarithm operation becomes

$$\log(s_f + n_f) = \log(n_f) + \log\left(1 + \frac{s_f}{n_f}\right), \tag{2}$$

where $(n_1, n_2, \ldots, n_F)^{\mathsf{T}}$ is the noise spectrum. The right hand side of (2) is evident from the Taylor series of $\log(x + y)$, and emphasises that CMN would remove the constant term $\log(n_f)$.

B. Properties of SNR features

It appears from the above analysis that, if we use CMN, the features that are presented to the ASR decoder are actually (a linear transform of) the logarithm of one plus the signal to noise ratio (SNR). This will happen even if the additive noise is simply the minimal background noise usually associated with clean recordings. It follows that we could try to calculate the SNR from the outset rather than calculate a spectral power measure and rely on CMN to produce the SNR. A-priori, such an approach has at least two appealing properties:

1) The flooring of the logarithm happens naturally. SNR values cannot fall below zero, so the argument of the logarithm is naturally floored at unity.
2) SNR is inherently independent of gain associated with microphones and pre-amplifiers.

We will show that SNR is also mathematically appealing.

The approach is analogous to that of Lathoud *et al.* [7]. The only difference is that Lathoud *et al.* explicitly floor the SNR using (in our present notation)

$$\max\left(1, \frac{s_f}{n_f}\right). \tag{3}$$

III. A MORE RIGOROUS ANALYSIS

In contrast to the previous section, which was left deliberately simplistic, we now present a more rigorous derivation of a SNR based feature. We begin by defining a Gaussian model of speech in noise, and proceed by showing that power spectral subtraction can be seen as a particular maximum-likelihood (ML) solution. We then derive ML and MAP estimators for the SNR.

A. Gaussian model

Let us assume that a DFT operation produces a vector, $\mathfrak{x}$, with complex components, $\mathfrak{x}_1, \mathfrak{x}_2, \ldots, \mathfrak{x}_F$, where the real and imaginary parts of each $\mathfrak{x}_f$ are i.i.d. normally distributed with zero mean and variance υ_f. That is,

$$f\left(\mathfrak{x}_f \mid \upsilon_f\right) = \frac{1}{\pi\upsilon_f} \exp\left(-\frac{|\mathfrak{x}_f|^2}{\upsilon_f}\right). \tag{4}$$

In the case where we distinguish two coloured noise signals, a background noise, $\mathfrak{n}$, and a signal of interest, $\mathfrak{s}$, typically speech, denote the noise variance as $\boldsymbol{\nu}$ and the speech variance as $\boldsymbol{\varsigma}$. In general, the background noise can be observed in isolation and modelled as

$$f\left(\mathfrak{n}_f \mid \nu_f\right) = \frac{1}{\pi\nu_f} \exp\left(-\frac{|\mathfrak{n}_f|^2}{\nu_f}\right). \tag{5}$$

The speech, however, cannot normally be observed in isolation. It is always added to noise. When both speech and additive noise are present the variances add, meaning that the total signal, $\mathfrak{t}_f = \mathfrak{s}_f + \mathfrak{n}_f$, can be modelled as

$$f\left(\mathfrak{t}_f \mid \varsigma_f, \nu_f\right) = \frac{1}{\pi(\varsigma_f + \nu_f)} \exp\left(-\frac{|\mathfrak{t}_f|^2}{\varsigma_f + \nu_f}\right). \tag{6}$$

The above model is the basis of the Wiener filter and of the widely used Ephraim-Malah speech enhancement technique [3]. The goal is usually formulated as requiring an estimate of $\mathfrak{s}_f$; this proceeds via estimation of ς_f.

We assume that an estimate, $\hat{\boldsymbol{\nu}}$, of $\boldsymbol{\nu}$ is available via solution of (5) during, for instance, non-speech segments of the signal.

Consider using (6) as a basis for estimation of the speech variance, $\boldsymbol{\varsigma}$. We drop the f subscript for simplicity. Bayes' theorem gives

$$f\left(\varsigma \mid \mathfrak{t}, \hat{\nu}\right) \propto f\left(\mathfrak{t} \mid \varsigma, \hat{\nu}\right) f\left(\varsigma\right). \tag{7}$$

If we assume a flat prior $f\left(\varsigma\right) \propto 1$, substituting (6) into (7), differentiating with respect to ς and equating to zero gives the well known maximum likelihood estimate,

$$\hat{\varsigma} = \max\left(|\mathfrak{t}|^2 - \hat{\nu}, 0\right). \tag{8}$$

This is known to provide a "reasonable" estimate of the speech variance, but always requires regularisation. In ASR, the ML estimate is known as power spectral subtraction. It is regularised by means of an over-subtraction factor, α, and a flooring factor, β:

$$\hat{\varsigma} = \max\left(|\mathfrak{t}|^2 - \alpha\hat{\nu}, \beta\hat{\nu}\right). \tag{9}$$

B. ML SNR estimate

The purpose of the above derivation is to show that a commonly used speech feature can be seen in a Bayesian sense as an estimate of the variance ς. We now follow the same procedure, but aim from the outset to estimate SNR. Define

$$\xi_f = \frac{\varsigma_f}{\nu_f}, \tag{10}$$

where ξ_f is exactly the *a-priori* SNR of McAulay and Malpass [9], popularised by Ephraim and Malah [3]. The f subscript indicates that the SNR is frequency dependent. Substituting $\varsigma_f = \xi_f\nu_f$ into (6),

$$f\left(\mathfrak{t}_f \mid \xi_f, \nu_f\right) = \frac{1}{\pi\nu_f(1 + \xi_f)} \exp\left(-\frac{|\mathfrak{t}_f|^2}{\nu_f(1 + \xi_f)}\right). \tag{11}$$

The subscript is dropped again hereafter for simplicity.

This time, the posterior is in terms of ξ,

$$f\left(\xi \mid \mathfrak{t}, \hat{\nu}\right) \propto f\left(\mathfrak{t} \mid \xi, \hat{\nu}\right) f\left(\xi\right). \tag{12}$$

Assuming a flat prior, substituting (11) into (12), differentiating and equating to zero,

$$\hat{\xi} = \max\left(\frac{|\mathfrak{t}|^2}{\hat{\nu}} - 1, 0\right). \tag{13}$$

It is shown in section IV that this result requires no further normalisation to work well. Further, notice that

$$\log(1+\hat{\xi}) = \log\left(\max\left[1, \frac{|\mathfrak{t}|^2}{\hat{\nu}}\right]\right), \tag{14}$$

which is the same form as (3). However, no ad-hoc spectral model is necessary.

We note that in the Decision Directed estimator of [3], the ML estimate of ξ of (13) is regularised using an estimate based on the previous spectral magnitude estimate. This is further explored by Cohen [10], and is used in a modified form in [4], [11]. Whilst these approaches are beyond the scope of the present study, our approach does not preclude using them.

C. Marginalisation over noise variance

Thus far we have assumed that an estimate, $\hat{\nu}$, of the noise variance is available. The form of (11), however, with multiplicative instead of additive terms in the denominators, allows marginalisation over the noise variance.

If we have N frames (spectral vectors) of noise, $\{\mathfrak{n}\}_N = \{\mathfrak{n}_1, \mathfrak{n}_2, \ldots, \mathfrak{n}_N\}$, that are observed in isolation, we can write

$$f(\nu_f \mid \{\mathfrak{n}\}_N) = \frac{\prod_{i=1}^{N} f(\mathfrak{n}_{i,f} \mid \nu_f)\, f(\nu_f)}{\int_0^\infty d\nu' \prod_{i=1}^{N} f\left(\mathfrak{n}_{i,f} \mid \nu'_f\right) f\left(\nu'_f\right)}, \tag{15}$$

where the products are over the likelihood terms, not the priors. Again, hereafter we drop subscripts for simplicity. The likelihood terms are exactly the form of equation (5), and we arbitrarily choose a non-informative prior $f(\nu) \propto \nu^{-1}$. Equation (15) then reduces to the inverse gamma distribution

$$f(\nu \mid \{\mathfrak{n}\}_N) = \frac{B^A}{\Gamma(A)} \nu^{-A-1} \exp\left(-\frac{B}{\nu}\right) \tag{16}$$

where

$$A = N, \qquad B = \sum_{i=1}^{N} |\mathfrak{n}_{i,f}|^2 . \tag{17}$$

The MAP solution, $\hat{\nu}$, of ν would be

$$\hat{\nu} = \frac{B}{A+1}, \tag{18}$$

however, we can use the distribution to marginalise over ν. Equation (12) becomes

$$f(\xi \mid \mathfrak{t}) \propto f(\xi) \int_0^\infty d\nu\, f(\mathfrak{t} \mid \xi, \nu)\, f(\nu \mid \{\mathfrak{n}\}_N). \tag{19}$$

Substituting (11) and (16) into (19), the forms are conjugate and the integral is just the normalising term from the inverse gamma distribution.

$$f(\xi \mid \mathfrak{t}) \propto f(\xi) \times \frac{B^A}{\Gamma(A)} \frac{\Gamma(A+1)}{\xi+1} \left(\frac{|\mathfrak{t}|^2 + (\xi+1)B}{\xi+1}\right)^{-(A+1)}. \tag{20}$$

D. Marginal ML estimate

If we assume a flat prior, $f(\xi) \propto 1$, as before, differentiating (20) and equating to zero gives

$$\hat{\xi} = \max\left(\frac{A|\mathfrak{t}|^2}{B} - 1, 0\right) \tag{21}$$

Curiously, equation (21) is basically the same as equation (13).

E. MAP estimate

Instead of using a flat (improper) prior for the speech variance, it is possible to use a proper prior representing real information. The prior distribution should allow (encourage, even) the SNR to be zero, but should discourage large values; greater than a few tens of decibels. Here we use the gamma distribution

$$f(\xi \mid \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \xi^{\alpha-1} \exp\left(-\frac{\xi}{\beta}\right). \tag{22}$$

Substituting this into (20), differentiating and equating to zero yields a cubic in ξ

$$a_3\xi^3 + a_2\xi^2 + a_1\xi + a_0 = 0, \tag{23}$$

with

$$\begin{aligned} a_3 &= -1, \\ a_2 &= -\beta + (\alpha-1)\beta - |\mathfrak{t}|^2/B - 2, \\ a_1 &= -\beta + \beta A|\mathfrak{t}|^2/B + (\alpha-1)\beta|\mathfrak{t}|^2/B \\ &\quad + 2(\alpha-1)\beta - |\mathfrak{t}|^2/B - 1, \\ a_0 &= (\alpha-1)\beta + (\alpha-1)\beta|\mathfrak{t}|^2/B, \end{aligned} \tag{24}$$

The cubic is readily solved using the cubic equation [12], and always has at least one real root. The root can, however, be negative, so the resulting $\hat{\xi}$ should be floored at zero.

To set the hyper-parameters, we find that simply constraining the expectation of the gamma distribution to be the average ML SNR of the current frame works satisfactorily,

$$E(\xi_f) = \alpha\beta_f \tag{25}$$

$$\beta_f = \frac{1}{\alpha} E(\xi_f) = \frac{1}{\alpha}\left[\frac{1}{F}\sum_{f=1}^{F} \frac{|\mathfrak{t}_f|^2}{\hat{\nu}_f} - 1\right], \tag{26}$$

and, empirically, $\alpha = 0.01$.

For illustration, figure 1 shows a histogram of SNR (actually $|\mathfrak{t}|^2/\hat{\nu}$) at 1000 Hz for the clean part of the aurora 2 training data. Also shown is a gamma distribution with $\alpha = 0.01$ and β set such that the expectation is 48dB, the approximate SNR of the clean aurora 2 data. The plot is in the log domain. Notice that the gamma distribution is basically flat (caused by α being close to 0), but falls rapidly for high values, i.e., it is largely uninformative but discourages high SNR.

We choose a gamma prior in this study for simplicity. Other authors (a recent example is [13]) have made persuasive cases for the speech prior being closer to a generalised gamma distribution. In ASR, the speech prior is often represented by a large Gaussian mixture [14].

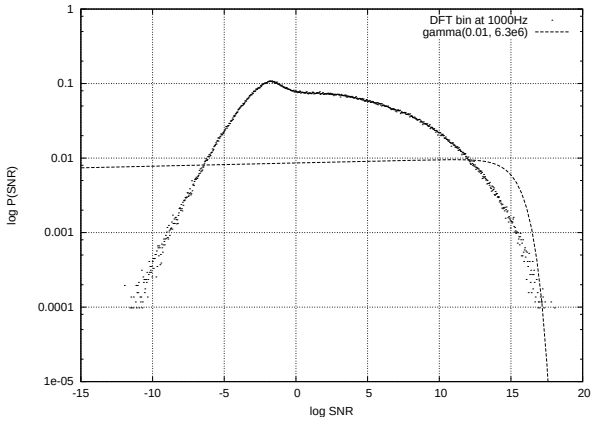

Fig. 1. Histogram of clean data at 1000 Hz and gamma distribution with $\alpha = 0.01$ and $\alpha\beta = 48$dB. The gamma distribution is largely flat (uninformative), but imposes an upper limit. log(SNR) would normally be floored at 0.

IV. Experiments

To allow comparison with [7] we present experimental results on aurora 2. The aurora 2 task [15] is a well known evaluation for noise compensation techniques. It is a simple digit recognition task with real noise artificially added in 5dB increments such that performance without noise compensation ranges from almost perfect to almost random. Both clean (uncorrupted) and multi-condition (additive noise corrupted) training sets are provided, along with three test sets:

- A Data corrupted with the same noise used in the (multi-condition) training.
- B As test set A, but using different noise.
- C A subset of the noises above, but additionally with a convolutional filter.

Aurora 2 does not distinguish evaluation and test data, so results may be biased towards this dataset and should be considered optimistic.

We used a simple "MFCC" front-end with a 256 point DFT every 10ms. The noise reduction techniques were applied in the power-spectral domain (129 bins), after which a filter bank of 23 mel-spaced triangular bins was applied. The usual logarithm and truncated DCT then produced 13 cepstral coefficients (including C0) plus first and second order delta coefficients. Where CMN and CVN were applied, the means and variances were calculated separately for the whole of each utterance.

The noise values were obtained using the low-energy envelope tracking method described in [5], but with a simplified correction factor from [16]: The 20 lowest energy samples in a sliding 100 frame (1 second) window were averaged, and multiplied by a correction factor, C. See section V-B for a discussion of this factor.

Complete results are shown in Figure 2. Each graph represents a full aurora evaluation with both multi-condition and clean training. The SNR of clean testing data was measured to be around 48dB, and is off the axis, but the result is shown as the first number in parentheses in the legend. The second number in the legend is the usual aurora figure of merit: the average of the scores from 0dB to 20dB.

Each graph in the left column represents use of CMN, whereas the right column represents use of CVN (implying CMN also). The four rows are, respectively, the value passed to the filter-bank being

1) The usual non-SNR (power spectral) features.
2) As 1, but with spectral subtraction.
$$\hat{\varsigma} = \max\left(|\mathfrak{t}|^2 - \alpha\hat{\nu}, \beta\hat{\nu}\right), \tag{27}$$
with $\alpha = 1$ and $\beta = 0.1$, found with a coarse grid search.
3) One plus the maximum likelihood estimate of SNR from the marginal distribution
$$\hat{\xi} + 1 = \max\left(\frac{A\,|\mathfrak{t}|^2}{B} - 1, 0\right) + 1, \tag{28}$$
$$= \max\left(\frac{A\,|\mathfrak{t}|^2}{B}, 1\right). \tag{29}$$
4) One plus the MAP estimate of the SNR with a gamma prior,
$$\hat{\xi} + 1, \tag{30}$$
where $\hat{\xi}$ is the solution of the cubic in (23) and (24).

We stress that these results are not state of the art for this database; the purpose is to compare techniques.

V. Discussion

A. Performance

The most significant result of these experiments is that the CVN results for the SNR features agree with, even exceed, those in [7]. This is despite the fact that no involved spectral model is used to distinguish the speech and noise. It seems that simply being able to track the background noise level with the low-energy envelope is enough.

The use of the simple gamma prior has a small benefit, but at the cost of an extra parameter and finding the solution to a cubic equation. Whilst this is not computationally onerous, it is doubtful whether it is worthwhile given the good performance of the much simpler ML solution. However, the spirit of the approach is important; it shows a principled way to incorporate prior information.

Spectral subtraction gives an improvement over the baseline, but does not respond to CVN. This is at odds with the results in [8], but in agreement with our own anecdotal evidence. This is a curious result since there is not a large theoretical difference between SNR features and spectral subtraction. The practical difference between the two is that SNR features normalise before the filter-bank, whereas CMN works after it. If we denote the filter-bank weights for a single bin by $w_1, w_2, \ldots$, the SNR features presented to the decoder are of the form
$$\log\left(1 + w_1\xi_1 + w_2\xi_2 + \cdots\right), \tag{31}$$

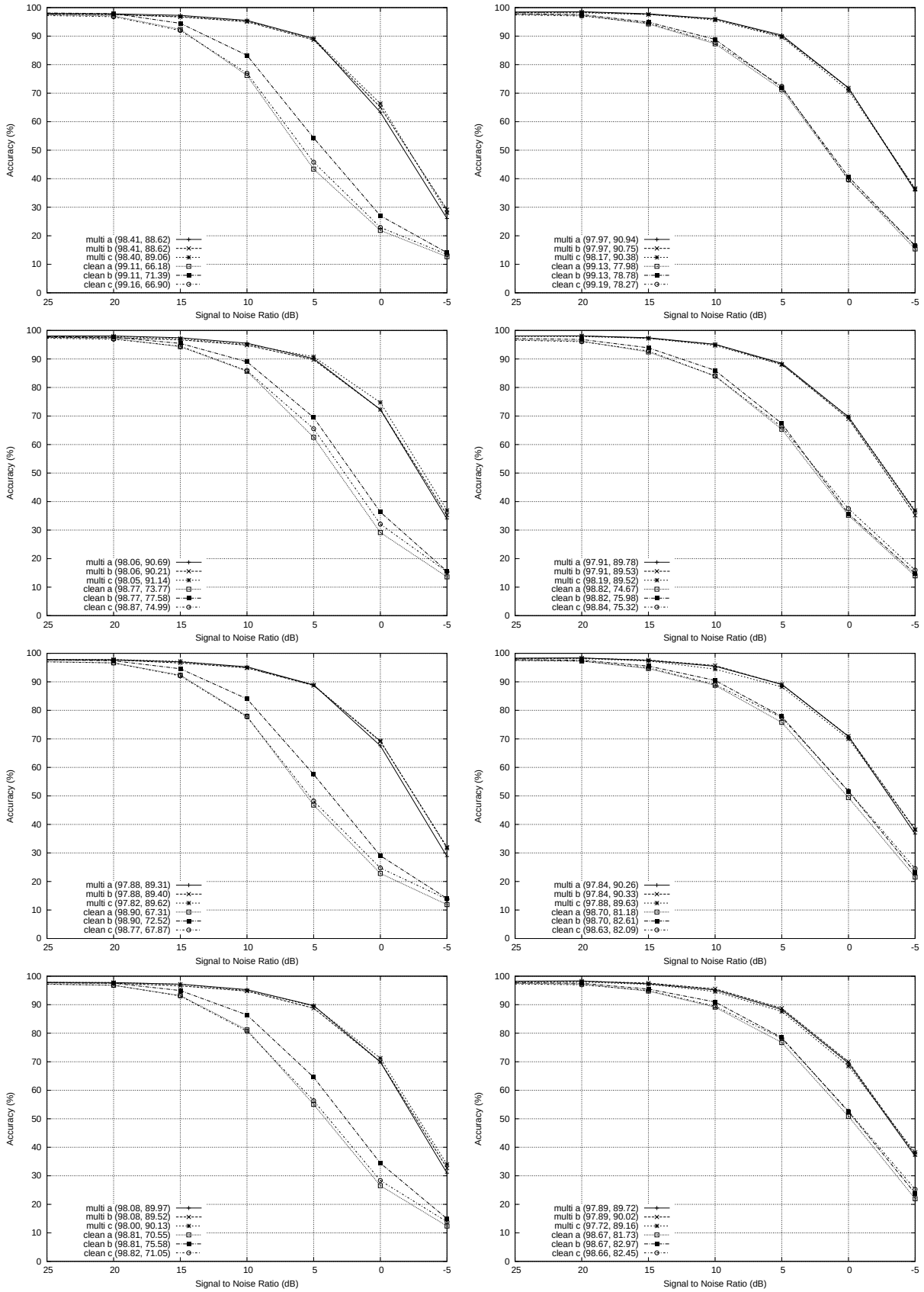

Fig. 2. Results. The left column is for CMN, right for CMN+CVN. The four rows top to bottom are respectively: No noise compensation, Spectral Subtraction, ML SNR and MAP SNR with gamma prior. See the text for details of the database.

whereas the spectral subtraction features are *closer* to the form

$$\log\left(1+\frac{w_1 s_1 + w_2 s_2 + \cdots}{w_1 n_1 + w_2 n_2 + \cdots}\right). \tag{32}$$

Given that $\log(a+b) \approx \log\max(a,b)$, we hypothesise that a large noise component anywhere in the band spanned by the given bin could dominate the latter expression. This in turn offers some explanation for the improved performance of SNR features. It remains a subject for further investigation.

B. Noise tracker correction factor

The low-energy envelope tracker normally requires correction as its estimate is biased too small. In [16], Lathoud *et al.* suggest that a multiplicative correction factor

$$C = \frac{1}{(1.5\gamma)^2}, \tag{33}$$

works well, where γ is the fraction of samples assumed to be noise. In our case, $\gamma = 0.2$ so $C = 11.1$. In fact, we found that, whilst this correction factor was necessary for the spectral subtraction approach, a value of $C = 1$ was better for SNR features (the results in Figure 2 are for these values).

It is tempting to conclude that SNR features do not need a correction factor. However, it is more likely that the noise tracker with $C = 1$ was producing noise estimates about 11 times too small, so the SNR estimates were 11 times too large. Writing the situation as

$$\log(1+11\xi) = \log(11) + \log\left(\frac{1}{11}+\xi\right), \tag{34}$$

it is clear that this corresponds to using a smaller floor in the logarithm. This floor is also very close to the one empirically found to work well as the parameter β in spectral subtraction.

The low-energy envelope is a noise floor rather than a noise estimate; it is intuitively reasonable that this floor is also the level below which speech and noise cannot be distinguished. We hypothesise that the optimal value of C in low-energy envelope tracking is the same as the optimal floor for SNR. Thus, when using SNR based features, these values cancel out giving a parameter-free feature. Proof that $C = 1$ is optimal for SNR features, however, will require a careful mathematical and experimental analysis.

VI. Conclusions

SNR features for ASR have several practical and mathematical advantages over the more usual spectral power features. The naive SNR estimate is actually the optimal estimate under a fairly rigorous Bayesian analysis, and the framework leaves room for further incorporation of prior information, as is common recently in ASR. SNR features perform well in noisy conditions, and outperform other features when combined with CVN. Prior information incorporated via a gamma prior distribution improves results still further, although the difference may not merit the extra complexity. In practice a different prior form, or one trained on real data ought to work better.

We have some evidence that the optimal correction factor used in low-energy envelope tracking cancels exactly the flooring used in the logarithm for SNR features, making SNR features almost parameter-free when noise is estimated in this manner.

VII. Acknowledgements

The author is grateful to Mathew Magimai-Doss for comments on the manuscript. This work was supported by the Swiss National Center of Competence in Research on Interactive Multi-modal Information Management. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

[1] O. Viikki and K. Laurila, "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization," in *Robust Speech Recognition for Unknown Communication Channels*. ISCA, April 1997, pp. 107–110, Pont-à-Mousson, France.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, pp. 113–120, April 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[4] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ETSI, ETSI Standard 202 050, 2002, v1.1.1.

[5] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Communication*, no. 34, pp. 141–158, 2001.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[7] G. Lathoud, M. Magimai-Doss, B. Mesot, and H. Bourlard, "Unsupervised spectral substraction for noise-robust ASR," in *Proceedings of the 2005 IEEE ASRU Workshop*, December 2005, San Juan, Puerto Rico.

[8] J. C. Segura, M. C. Benítez, A. de la Torre, and A. J. Rubio, "Feature extraction combining spectral noise reduction and cepstral histogram equalisation for robust ASR," in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 225–228.

[9] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, April 1980.

[10] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, September 2005.

[11] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, Montreal, Canada, May 2004, pp. 289–292.

[12] "Cubic function," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Cubic_function

[13] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1741–1752, August 2007.

[14] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Atlanta, US, May 1996, pp. 733–736.

[15] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millenium"*, September 2000, Paris, France.

[16] G. Lathoud, M. Magimai-Doss, and H. Bourlard, "Channel normalization for unsupervised spectral subtraction," Idiap Research Institute, IDIAP-RR 06-09, February 2006. [Online]. Available: http://publications.idiap.ch