View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

IDIAP RESEARCH REPORT

CEPSTRAL NORMALISATION AND THE SIGNAL TO NOISE RATIO SPECTRUM IN AUTOMATIC SPEECH RECOGNITION.

Philip N. Garner

Idiap-RR-15-2011

MAY 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition.

Philip N. Garner

May 31, 2011

Abstract

Cepstral normalisation in automatic speech recognition is investigated in the context of robustness to additive noise. In this paper, it is argued that such normalisation leads naturally to a speech feature based on signal to noise ratio rather than absolute energy (or power). Explicit calculation of this *SNR-cepstrum* by means of a noise estimate is shown to have theoretical and practical advantages over the usual (energy based) cepstrum. The relationship between the SNR-cepstrum and the articulation index, known in psycho-acoustics, is discussed. Experiments are presented suggesting that the combination of the SNR-cepstrum with the well known perceptual linear prediction method can be beneficial in noisy environments.

1 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. This paper concentrates 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.

The present work stems from the practical experience that it is very difficult to improve upon cepstral normalisation techniques for noise robustness. Cepstral mean normalisation (CMN) (Furui, 1981) is a well established technique that compensates, in a theoretically sound way, for convolutional noise. It is based on the persuasive observation that a linear channel distortion becomes a constant offset in the cepstral domain. More heuristically, CMN also affords some robustness to additive noise. Cepstral variance normalisation (CVN) (Viikki and Laurila, 1997, 1998) generally results in very good noise robustness, but the reason for this is not well understood.

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 (1979) is well established. In speech enhancement, much work is based on the technique of Ephraim and Malah (1984). Both these techniques have influenced the design of the ETSI (2002) standard ASR front-end. However, at least in a batch mode of operation, and certainly combined with multi-condition training, CMN combined with CVN can exceed the performance of all these techniques.

In this paper, building on previous work, the theoretical effect of CMN and CVN in additive noise is studied. It is shown that the use of CMN implies that the features presented to an ASR decoder are in fact measures of (log) signal to noise ratio (SNR) rather than (log) energy. Based on this observation, a SNR feature is derived formally, the derivation providing both theoretical and practical advantages over the equivalent for energy based features.

The SNR-cepstrum is then placed in context amongst other techniques, emphasising that there is a great deal of commonality between noise robustness in ASR, speech enhancement and indeed the workings of the inner ear.

The paper is split roughly into two parts. Sections 1 to 4 are largely theoretical, expanding previous work to give a thorough basis for the SNR-cepstrum. Sections 5 to 7 proceed to evaluate the SNR-cepstrum in the context of the linear predictive features that are common in modern ASR systems.

2 Background

In a simplistic, but informative, view of an ASR front-end, an acoustic signal is Fourier transformed to give a vector of spectral coefficients $(s_1, s_2, \ldots, s_F)^T$. After a linear transform (filter-bank) implementing a non-linear frequency warp, the cepstrum is calculated. The cepstrum involves a logarithm followed by another linear transform (DCT).

2.1 Convolutional noise

Although only one is normally considered, note that two types of convolutional noise can be distinguished:

1. A *source* noise, $\mathbf{g} = (g_1, g_2, \ldots, g_F)^\mathsf{T}$, associated only with the speech signal. This can be thought of as being representative of a speaker.
2. A *channel* noise, $\mathbf{h} = (h_1, h_2, \ldots, h_F)^\mathsf{T}$, associated with the microphone and transmission channel.

In the presence of convolutional noise, which is multiplicative in the frequency domain, the logarithm for each frequency bin, $f \in \{1, 2, \ldots, F\}$, becomes

$$\log(h_f g_f s_f) = \log(h_f) + \log(g_f) + \log(s_f), \quad (1)$$

where $\log(s_f)$ varies, and $\log(h_f)$ is constant over time. $\log(g_f)$ is taken to represent the component of the speech that is constant over time, being some characteristic of the speaker.

It follows from equation 1 that, if $\log(s_f)$ can be assumed to have zero mean, the noise terms can be removed by subtracting the long term average of the log-spectrum. This is achieved by cepstral mean normalisation (Furui, 1981, although the technique has been attributed to Atal even earlier) or by the RASTA processing of Hermansky and Morgan (1994). Note also that, when the filter-bank is considered, the above holds if the h_f and g_f are assumed constant within a given filter-bank bin.

2.2 Additive noise

When additive noise is also present, typically it is assumed to remain additive after the Fourier transform. In this sense, the logarithm operation becomes

$$\log(h_f g_f s_f + h_f n_f) = \log(h_f) + \log(g_f s_f + n_f). \quad (2)$$

where $(n_1, n_2, \ldots, n_F)^\mathsf{T}$ is the noise spectrum. From equation 2, it appears that CMN and the like cannot work in significant additive noise unless the additive noise is removed first. To this end, there is a large body of work focusing on additive noise removal. In ASR, the spectral subtraction approach of Boll (1979) was further developed by, for instance, Van Compernolle (1989), and is well established. It is often used as a means to derive a Wiener filter. In speech enhancement, much work is based on the technique of Ephraim and Malah (1984).

The state of the art in additive noise robustness is probably in the body of work based on the additive model of Acero and Stern (1990); Acero (1990), and the vector Taylor series approach of Moreno et al. (1996); Moreno (1996). Such techniques are characterised by a large Gaussian mixture prior on the speech signal, a recent exemplar being Li et al. (2007). It is not the goal of the present paper to approach the performance of such techniques. Rather, a building block is presented that could be used in combination with these techniques.

2.3 SNR features

The logarithm of a sum can be written

$$\begin{aligned} \log(x + a) &= \log(a) + \frac{x}{a} - \frac{x^2}{2a^2} + \frac{x^3}{3a^3} \cdots \\ &= \log(a) + \log\left(1 + \frac{x}{a}\right). \end{aligned} \quad (3)$$

Although the relationship is clear without the series expansion, the latter emphasises that the term $\log(a)$ is the component that is independent of x. This in turn suggests that equation 2 might better be written

$$\log(h_f g_f s_f + h_f n_f) = \log(h_f n_f) + \log\left(1 + \frac{g_f s_f}{n_f}\right), \quad (4)$$

emphasising that CMN would actually remove the constant term $\log(h_f n_f)$, or its mean if either h_f or n_f were non-deterministic.

It appears from the above analysis that, if CMN is used, 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 one 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 three appealing properties:

1. The flooring of the logarithm happens naturally. The SNR (expressed as a power ratio) cannot fall below zero, so the argument of the logarithm is naturally floored at unity, and the logarithm is hence positive.
2. SNR is inherently independent of $\mathbf{h}$, the convolutional noise associated with microphones and the gain associated with pre-amplifiers.
3. If applied before the filter bank, the assumption that h_f remains constant over the range of the filter bin is no longer required.

It turns out that SNR is also mathematically appealing.

Notice that, whilst the channel noise, h_f, is cancelled by taking the SNR, the source noise, g_f, is still present. However, for high SNR it will be removed by CMN. It follows that the SNR is not a replacement for CMN in its speaker normalisation sense. It also suggests that direct comparison of SNR based features with CMN would not be fair.

3 The SNR spectrum

In contrast to the previous section, which was left deliberately simplistic, a more rigorous derivation of a SNR based feature is now presented. After defining a Gaussian model of speech in noise, the derivation proceeds

by showing that power spectral subtraction can be seen as a particular maximum-likelihood (ML) solution. Two ML estimators for the SNR are then derived.

3.1 Gaussian model

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 Gaussian, independent and identically distributed (i.i.d.) with zero mean and variance υ_f. That is,

$$p(\mathfrak{x}_f \mid \upsilon_f) = \frac{1}{\pi\upsilon_f} \exp\left(-\frac{|\mathfrak{x}_f|^2}{\upsilon_f}\right). \tag{5}$$

In the case where two coloured noise signals are distinguished, 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 $\boldsymbol{\sigma}$. In general, the background noise can be observed in isolation and modelled as

$$p(\mathfrak{n}_f \mid \nu_f) = \frac{1}{\pi\nu_f} \exp\left(-\frac{|\mathfrak{n}_f|^2}{\nu_f}\right). \tag{6}$$

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

$$p(\mathfrak{t}_f \mid \sigma_f, \nu_f) = \frac{1}{\pi(\sigma_f + \nu_f)} \exp\left(-\frac{|\mathfrak{t}_f|^2}{\sigma_f + \nu_f}\right). \tag{7}$$

Although neither the Gaussian nor i.i.d. assumptions are likely to be true in practice, the above model is the basis of the Wiener filter and of the widely used Ephraim and Malah (1984) speech enhancement technique. The goal is usually formulated as requiring an estimate of $\mathfrak{s}_f$. However, it is first necessary to find an estimate of σ_f.

3.2 Variance as an ASR feature

The well known maximum likelihood estimate of σ_f is instructive in determining the right approach for the definition and estimation of SNR. It proceeds as follows, where the f subscript is dropped for simplicity: Assume that an estimate, $\hat{\nu}$, of $\boldsymbol{\nu}$ is available via solution of (6) during, for instance, non-speech segments of the signal. The estimate of the speech variance, $\boldsymbol{\sigma}$, then follows from Bayes' theorem,

$$p(\sigma \mid \mathfrak{t}, \hat{\nu}) \propto p(\mathfrak{t} \mid \sigma, \hat{\nu})\, p(\sigma \mid \hat{\nu}). \tag{8}$$

Assuming $p(\sigma \mid \hat{\nu}) = p(\sigma)\, p(\hat{\nu})$ and a flat prior $p(\sigma) \propto 1$, substituting (7) into (8), differentiating with respect to σ and equating to zero gives the ML estimate,

$$\hat{\sigma} = \max\left(|\mathfrak{t}|^2 - \hat{\nu}, 0\right). \tag{9}$$

Notice that, in ASR at least, this is simply power spectral subtraction. More generally, it is known to provide a "reasonable" estimate of the speech variance, but always requires regularisation. In ASR, it is regularised by means of an over-subtraction factor, α, and a flooring factor, β:

$$\hat{\sigma} = \max\left(|\mathfrak{t}|^2 - \alpha\hat{\nu}, \beta\hat{\nu}\right), \tag{10}$$

as in Van Compernolle (1989).

The above derivation shows that a commonly used speech feature can be seen in a Bayesian sense as an estimate of the variance σ. This interpretation is reinforced when convolutional noise is considered. Making the substitution $\mathfrak{y}_f = \sqrt{h_f}\mathfrak{x}_f$ in equation 5, the Jacobian determinant is h_f^{-1}, so

$$p(\mathfrak{y}_f \mid h_f, \upsilon_f) = \frac{1}{\pi h_f \upsilon_f} \exp\left(-\frac{|\mathfrak{y}_f|^2}{h_f \upsilon_f}\right), \tag{11}$$

i.e., the convolutional term multiplies the variance, exactly as in the simplistic model of section 2.

The above implies that estimation for the purposes of ASR can focus on the variance, σ, rather than the (uncorrupted) observation, $\mathfrak{s}$, as in enhancement.

3.3 ML SNR estimate

Motivated by the term of interest being the variance, define the SNR as

$$\xi_f = \frac{\sigma_f}{\nu_f}, \tag{12}$$

The f subscript indicates that the SNR is frequency dependent. Substituting $\sigma_f = \xi_f \nu_f$ into (7),

$$p(\mathfrak{t}_f \mid \xi_f, \nu_f) = \frac{1}{\pi\nu_f(1+\xi_f)} \exp\left(-\frac{|\mathfrak{t}_f|^2}{\nu_f(1+\xi_f)}\right). \tag{13}$$

The subscript is dropped again hereafter for simplicity.

This time, the posterior is in terms of ξ,

$$p(\xi \mid \mathfrak{t}, \hat{\nu}) \propto p(\mathfrak{t} \mid \xi, \hat{\nu})\, p(\xi \mid \hat{\nu}). \tag{14}$$

Assuming a flat prior, substituting (13) into (14), differentiating and equating to zero,

$$\hat{\xi} = \max\left(\frac{|\mathfrak{t}|^2}{\hat{\nu}} - 1, 0\right). \tag{15}$$

3.4 Marginalisation over noise variance

Thus far it has been assumed that an estimate, $\hat{\nu}$, of the noise variance is available. In a Bayesian sense, however, the noise is a nuisance variable, the correct approach being to marginalise over it. In the case of variance estimation, such marginalisation is not easily tractable. By contrast, the form of (13), with multiplicative instead of

additive terms in the denominators, presents no major difficulty for marginalisation.

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

$$p(\nu_f \mid \{\mathbf{n}\}_N) = \frac{\prod_{i=1}^{N} p(n_{i,f} \mid \nu_f)\, p(\nu_f)}{\int_0^\infty d\nu' \prod_{i=1}^{N} p(n_{i,f} \mid \nu'_f)\, p(\nu'_f)}, \quad (16)$$

where the products are over the likelihood terms, not the priors. Again, hereafter subscripts are dropped for simplicity. The likelihood terms are exactly the form of equation (6), and a non-informative prior, $p(\nu) \propto \nu^{-1}$, is arbitrarily chosen. Equation (16) then reduces to the inverse gamma distribution

$$p(\nu \mid \{\mathbf{n}\}_N) = \frac{B^A}{\Gamma(A)} \nu^{-A-1} \exp\left(-\frac{B}{\nu}\right) \quad (17)$$

where

$$A = N, \qquad B = \sum_{i=1}^{N} |n_{i,f}|^2. \quad (18)$$

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

$$\hat{\nu} = \frac{B}{A+1}, \quad (19)$$

however, the distribution can be used to marginalise over ν. Assuming the prior on SNR is independent of the noise estimate, equation (14) becomes

$$p(\xi \mid t) \propto p(\xi) \int_0^\infty d\nu\, p(t \mid \xi, \nu)\, p(\nu \mid \{\mathbf{n}\}_N). \quad (20)$$

Substituting (13) and (17) into (20), the forms are conjugate and the integral is just the normalising term from the inverse gamma distribution.

$$p(\xi \mid t) \propto p(\xi) \times \frac{B^A}{\Gamma(A)} \frac{\Gamma(A+1)}{\xi+1} \left(\frac{|t|^2 + (\xi+1)B}{\xi+1}\right)^{-(A+1)}. \quad (21)$$

If a flat prior, $p(\xi) \propto 1$, is assumed as before, differentiating (21) and equating to zero gives a marginal ML estimate:

$$\hat{\xi} = \max\left(\frac{A|t|^2}{B} - 1, 0\right) \quad (22)$$

Curiously, equation (22) is basically the same as equation (15). It was shown by Garner (2009) that this result requires no further regularisation to work well.

Hereafter, the SNR vector, $\boldsymbol{\xi}$, is referred to as the SNR-spectrum. This leads to the resulting cepstrum being called the SNR-cepstrum.

4 Context

Whilst the above derivation is novel to the knowledge of the author, the SNR-spectrum is by no means a new concept. Rather, it draws together several loosely related topics.

4.1 Enhancement

ξ is exactly the *a-priori* SNR of McAulay and Malpass (1980), popularised by Ephraim and Malah (1984). In enhancement, this measure is used as an intermediate result in the reconstruction of an enhanced spectrum. The Wiener filter can be defined in terms of the SNR:

$$w = \frac{\xi}{\xi+1}. \quad (23)$$

In the decision directed estimator of Ephraim and Malah (1984), the ML estimate of ξ of (15) is regularised using an estimate based on the previous spectral magnitude estimate. This is further explored by Cohen (2005), and is used in a modified form in ETSI (2002); Plapous et al. (2004). Whilst these approaches are beyond the scope of the present study, the proposed approach does not preclude using them.

4.2 Automatic speech recognition

Lathoud et al. (2005) present an ad-hoc model allowing a signal to be described in terms of noise and speech spectra. Those authors perform what they refer to as "Unsupervised" spectral subtraction. In fact, they explicitly floor the SNR using (in the present notation)

$$\max\left(1, \frac{s_f}{n_f}\right). \quad (24)$$

Notice that

$$\log(1+\hat{\xi}) = \log\left(\max\left[1, \frac{|t|^2}{\hat{\nu}}\right]\right), \quad (25)$$

which is the same form as (24). However, no ad-hoc spectral model is necessary. It was shown by Garner (2009) that this formulation can actually exceed the performance reported by Lathoud et al. (2005).

The terminology raises an interesting issue: in the context of CMN, there is little difference between using the SNR-spectrum, and spectral subtraction. This is explored below in section 6.2.

4.3 Relationship with articulation index

Allen (1994) describes earlier work by Fletcher analysing the probable workings of the inner ear. In particular, Allen states that Fletcher's experiments suggest that the cochlea is sensitive to SNR:

> The signal to noise ratio of each cochlear inner hair cell signal is important to the formation of the feature channels since [the channel error] is known to depend directly on these SNRs rather than on spectral energy.

Later, Allen (2005) defines the articulation index (AI) as

$$\mathrm{AI}_k = \min\left(\frac{1}{3}\log_{10}(1 + c^2 \mathrm{snr}_k^2), 1\right). \qquad (26)$$

The AI is lower bounded at 0 by the logarithm, and upper bounded at 1 by a heuristic 30dB dynamic range of speech.

Notice that the AI has the same form, except for linear transformation, as the speech feature described above that arises from CMN. This in turn is known to work well in ASR. These two derivations are totally independent. It follows that, under CMN, the feature being presented to an ASR decoder is the AI, just as in the human ear.

In fact, the AI has been used directly as an ASR feature by Lobdell et al. (2008). The approach of those authors was to use the AI specifically to mimic the function of the ear. In this sense, the present approach is complementary, driven more mathematically than perceptually.

4.4 Noise tracker

In order to obtain a noise estimate, Garner (2009) used the low-energy envelope tracker advocated by Lathoud et al. (2006), based on Ris and Dupont (2001) and Martin (2001). The low-energy envelope tracker normally requires correction as its estimate is biased too small. Lathoud et al. (2006) suggest that a multiplicative correction factor

$$C = \frac{1}{(1.5\gamma)^2}, \qquad (27)$$

works well, where γ is the fraction of samples assumed to be noise. However, Garner (2009) found that a value of $C = 1$ was better for the SNR-cepstrum, rather than the $C \approx 11$ that would be implied from equation 27. This in turn implied that the feature being presented to the decoder was closer to

$$\log(1 + 11\xi) = \log(11) + \log\left(\frac{1}{11} + \xi\right). \qquad (28)$$

The right hand side of equation 28 implies that this corresponds to using a smaller floor in the logarithm. Further, it is close to the one empirically found to work well as the parameter β in spectral subtraction. However, the left hand side of equation 28 suggests a relationship with the AI: Allen (2005) states that the value c from equation 26 should be around 2. The square is certainly the same order of magnitude as the 11 that occurs empirically in the results of Garner (2009). C is based on noise minima and c is based on speech maxima; whatever the actual value of these constants, the present approach is unable to distinguish them. However, that they appear to cancel each other out suggests they have the same origin.

4.5 Cepstral variance normalisation

Whilst cepstral variance normalisation (CVN) is known to provide noise robustness (Viikki and Laurila, 1997, 1998), the justification for this is normally attributed to a heuristic and brute force shift of the observation PDF towards that of the model. This heuristic is used to good advantage in histogram normalisation (Segura et al., 2002; de la Torre et al., 2005). In the context of the SNR-spectrum, however, the concept of CVN is far more tangible: It is normalising SNR dynamic range.

As an aside, it follows that it may be possible to normalise for SNR at some other point in the processing chain. This has been investigated by the author without success. An obvious tentative conclusion is that the removal of the source noise, $\mathbf{g}$, via CMN is important beforehand.

4.6 Summary

The SNR-spectrum arises as a natural consequence of doing CMN on ASR features. CVN then takes on a physical interpretation as normalisation of the SNR dynamic range in dB. If defined more formally as the ratio of speech and noise variances, the intuitive estimator of SNR is also the marginal ML estimator under Gaussian noise.

The SNR-cepstrum appears to be exactly (differing only by linear transform) the AI of Fletcher as defined by Allen, suggesting a close relationship with the sensory mechanisms in the cochlea. Calculating the SNR-cepstrum as suggested both by the cochlea and practical computation leads to better noise robustness at low SNR.

5 Experiments

5.1 Previous results

Garner (2009) presented results showing that SNR based MFCC (mel frequency cepstral coefficients) features were more noise robust than the usual energy based features on the aurora 2 database. The aurora 2 task (Hirsch and Pearce, 2000) is a well known evaluation for noise compensation techniques. It is a simple English digit recognition task with real noise artificially added in 5 dB 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.

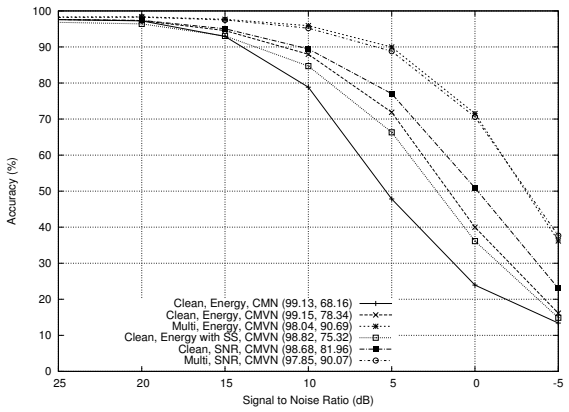

Figure 1: A summary of previous aurora 2 results for MFCC features. See the text for a description.

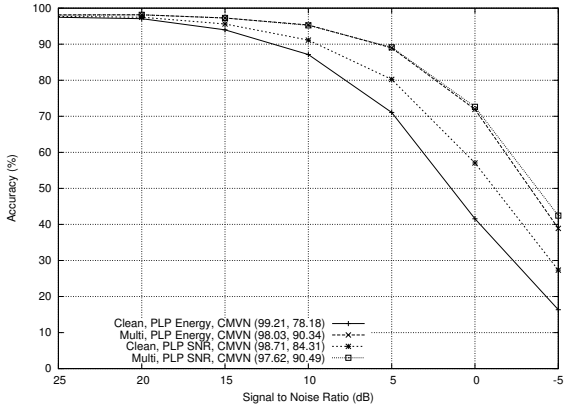

Figure 2: PLP results on aurora 2 database.

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 data-set and should be considered optimistic. It should also be stressed that the results in this paper are not state of the art for this database; the purpose is to compare techniques.

Aurora 2 is very useful for optimisation and evaluation of front-ends; this is because it runs quickly and has a thorough test set. However, several criticisms can be levelled at aurora 2:

1. It is real noise, but added artificially. This assumes that the additive noise assumption is exact, and ignores effects associated with the fact that speakers will modify their voices to compensate for noise presence.

2. It is digits, hence with a limited grammar and incomplete phonetic coverage.

There is also a somewhat intangible feeling in the community that aurora 2 results are often not reflected in real world systems.

The results from Garner (2009) are summarised in figure 1. Each graph represents a full aurora 2 evaluation for either multi-condition or clean training. As the results for the different test sets (A, B and C) are virtually indistinguishable when CMN is used, each curve is the average of the three sets. The SNR of clean testing data was measured to be around 48 dB, 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 2 figure of merit: the average of the scores from 0 dB to 20 dB. Both numbers are averaged over the three test sets.

The first curve in figure 1 shows an MFCC baseline using CMN in clean (mismatched) training conditions. The following two curves show the benefits of using CVN too (CMVN: cepstral mean and variance normalisation), and of multi-condition (matched) training. The next curve shows that spectral subtraction cannot improve on CVN, whilst the penultimate curve shows that the SNR-cepstrum can further improve on CVN in mismatched conditions. The final curve shows that the SNR-cepstrum does not afford any further improvement in matched conditions. In fact, all techniques perform very similarly under multi-condition training.

Notice that, whilst the aurora 2 figure of merit is higher for the SNR based features, it is mainly gained from improvements below about 15 dB SNR. In cleaner conditions, the usual energy based features perform better. It seems reasonable to attribute this difference to the noise tracker. Certainly the noise tracker is imperfect, and it is the only major difference between the two techniques at high SNR.

5.2 Hypotheses

In the present investigation, two hypotheses are under test:

1. State of the art systems often use linear prediction features as alternatives to the MFCCs used in previous work. Do such features also benefit from the use of SNR based features?

2. The previous experiments were limited to the scope of aurora 2. Do the benefits of SNR based features transcend the restrictions of the this database?

5.3 Perceptual linear prediction

Linear prediction (LP) is a common speech analysis method that represents speech using an all pole model (Makhoul, 1975). In the context of ASR, it is used to smooth a spectrum based on the fact that the signal originates from a vocal tract.

LP is normally used in ASR in the form of the perceptual linear prediction (PLP) of Hermansky (1990). PLP

modifies the auto-correlation calculation in the first stage of the LP calculation as follows:

1. The power spectrum is binned into critical bands separated according to the bark scale.
2. The bands are weighted according to an equal loudness criterion.
3. The bands are compressed by a cube root representing the power law of human hearing.

PLP has become quite widely used in state of the art ASR systems, e.g., the AMIDA system of Hain et al. (2010). In this sense, it merits investigation in the SNR-spectrum framework.

Whilst LP has a rigorous mathematical underpinning, PLP is more a set of heuristics. That is, the spectral warping is not derived as such, it is introduced in an ad-hoc, but intuitively reasonable manner. Using the same intuition, PLP cepstra can be calculated based on SNR rather than energy. If PLP is seen as simply a smoothing operation, it is reasonable to assume that the same smoothing can be applied to the SNR spectrum rather than the power (energy) spectrum.

5.4 Method

Features in the spirit of PLP were extracted using the Tracter toolkit (Garner and Dines, 2010). That is, preemphasis was used in lieu of an equal loudness weighting, then a 256 point DFT was performed every 10ms. The power spectrum of 129 bins was applied to a filter bank of 32 mel-spaced triangular bins (rather than bark spaced trapezoidal bins). The filter bank was cube root compressed (initially), then the usual DCT and LP recursions yielded 13 cepstral coefficients (including C0) plus first and second order delta coefficients. Cepstral means and variances were calculated separately for the whole of each utterance; all new results in this paper use both CMN and CVN.

The SNR based PLP features were extracted as above, except using one plus the ML estimate of the SNR as described in section 3.4. The LP calculation was as above, except that no cube root compression was employed. This was found to improve performance significantly, and is discussed later in section 6.3.

Following Garner (2009), the noise values were obtained using the low-energy envelope tracking method described by Ris and Dupont (2001), but with a simplified correction factor from Lathoud et al. (2006): The 20 lowest energy samples in a sliding 100 frame (1 second) window were averaged, but not multiplied by any correction factor.

5.5 Aurora 2 results

Results are shown in figure 2. The energy based PLP features perform similarly to the energy based MFCC features. However, the improvement for SNR based features is considerably more than that for MFCCs in the mismatched (clean training) case. This is encouraging; it strongly suggests not only that the SNR spectrum is applicable to PLP features, but that it is more suited to PLP features than to MFCCs.

5.6 Aurora 3 and 4 results

Aurora 3 and 4 go some way to combat the criticisms that are often levelled at aurora 2.

Aurora 3 is a digit subset of SpeechDat-Car; that is, a similar task to aurora 2 but uttered in real noise. The noise is various driving conditions of a car. Several languages are available; the present experiments are performed on the German (Netsch, 2001) and Danish (Lindberg, 2001) versions. As with aurora 2, a standardised train and test harness is provided using HTK. However, as the noise conditions are real, only three conditions are defined:

wm is *well-matched*; a mixture of all conditions and microphones for both training and testing.

mm is *mid-mismatch*; training with quiet and low noise data on a hands free microphone, testing on high noise data from the same microphone.

hm is *high-mismatch*; training in all conditions on a close talking microphone, testing in low and high noise on a hands free microphone.

No SNR information is immediately available for the Danish database. However, Netsch (2001) gives SNR distributions for the various microphones and conditions. The close talking microphone averages around 20 dB, and the hands free microphones averages around 5-10 dB; however all conditions spread 10 dB either side of the average. Given these broad measurements, and comparing with aurora 2 results, a-priori it may be expected that SNR features may not afford any improvement on the wm and mm conditions. However, an improvement is expected for the hm condition; although perhaps not as much as in aurora 2 as the mismatch is not as large.

Results are shown in figure 3. Contrary to expectations, there is a small improvement across the board, except for the Danish matched conditions. As expected, however, the improvement is most significant for the highest mismatch.

Aurora 4 is a noisy version of the well known wall street journal (WSJ) based SI-84 task. Aurora 4 goes back to using real noise artificially added to otherwise undistorted speech, but is large vocabulary (5000 words), hence covering the phone set thoroughly. As in aurora 2, both clean and multi-condition training sets are defined. However, rather than define tests at particular SNRs, 14 individual enumerated tests are specified; these are summarised in table 1.

Microphone	Clean	Noise added between 5 dB and 15 dB					
		Car	Babble	Restaurant	Street	Airport	Train
Sennheiser	1	2	3	4	5	6	7
Second	8	9	10	11	12	13	14

Table 1: Test set composition for aurora 4.

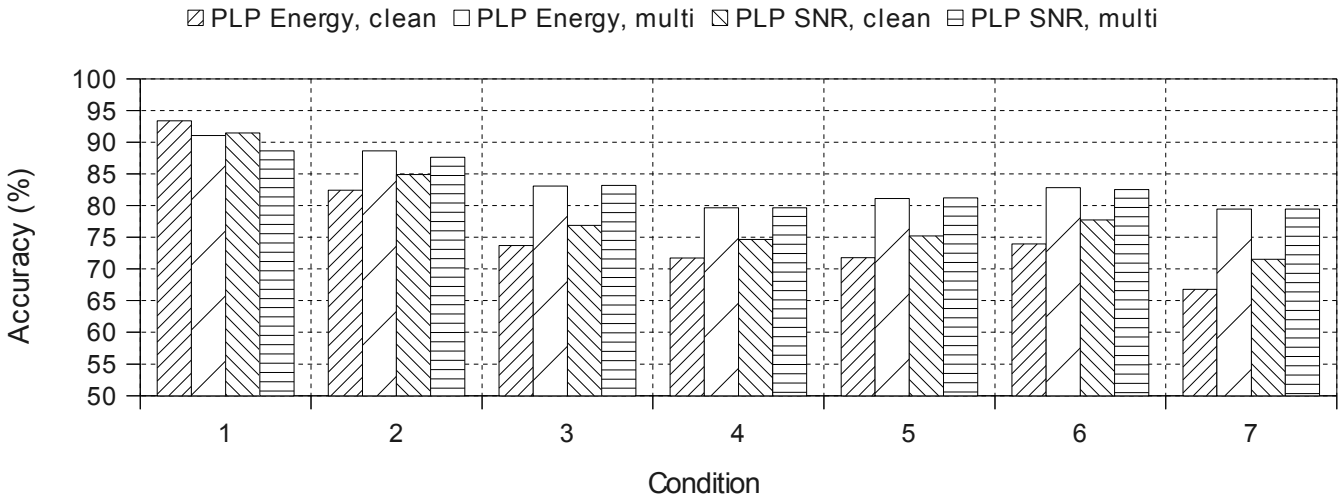

Figure 4: PLP results on aurora 4 database — Sennheiser microphone.

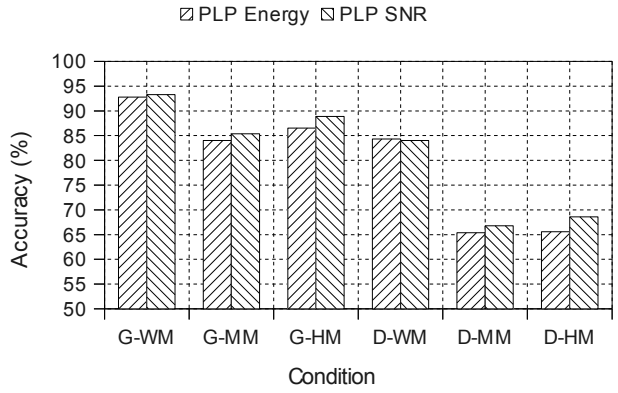

Figure 3: PLP results on aurora 3 database. The G and D prefixes refer to German and Danish respectively.

Although a test harness was made available by Parihar et al. (2004), other authors have written their own (e.g., Au Yeung and Siu, 2004). In the present experiments, a scheme in the spirit of that of Parihar et al. (2004), but using HTK, was used.

To better reflect a typical WSJ system, the 16 kHz data were used with a 400 point DFT and 40 bank mel filter. Other parameters were as in the 8 kHz experiments. Results are shown in figure 4 (Sennheiser microphone) and figure 5 (second microphone). A priori, from the aurora 2 result, one would not expect the multi-condition results to vary much between SNR and energy based PLPs. The added noise is in the range 5-15 dB, however, which is within the range in which SNR features have been shown to afford an improvement. In this sense, the clean training results should be better for SNR based PLPs.

In practice, the a-priori expectations are borne out quite well.

5.7 Rich text

The SNR-cepstrum was briefly evaluated in the context of meeting room recognition. The baseline was the AMIDA RT06 system of Hain et al. (2006). Only the first pass was evaluated, and only the IHM (individual headset microphone). At an early stage, it was clear that the results from the SNR-cepstrum were no better than those from the baseline, and further experiments were abandoned.

In fact, this result is broadly what would be expected a-priori given the aurora 2 results. The training and test condition are matched, and the SNR is quite high; perhaps better than the notional 15 dB threshold.

5.8 Experimental conclusions

The hypotheses are hence proven:

1. PLP features appear to benefit from SNR spectra in the same way as MFCC have been shown to do. At least on aurora 2, the results are better than for MFCCs.

2. Predictions made on the basis of aurora 2 results carry over to real noisy data, and to a large vocabulary system.

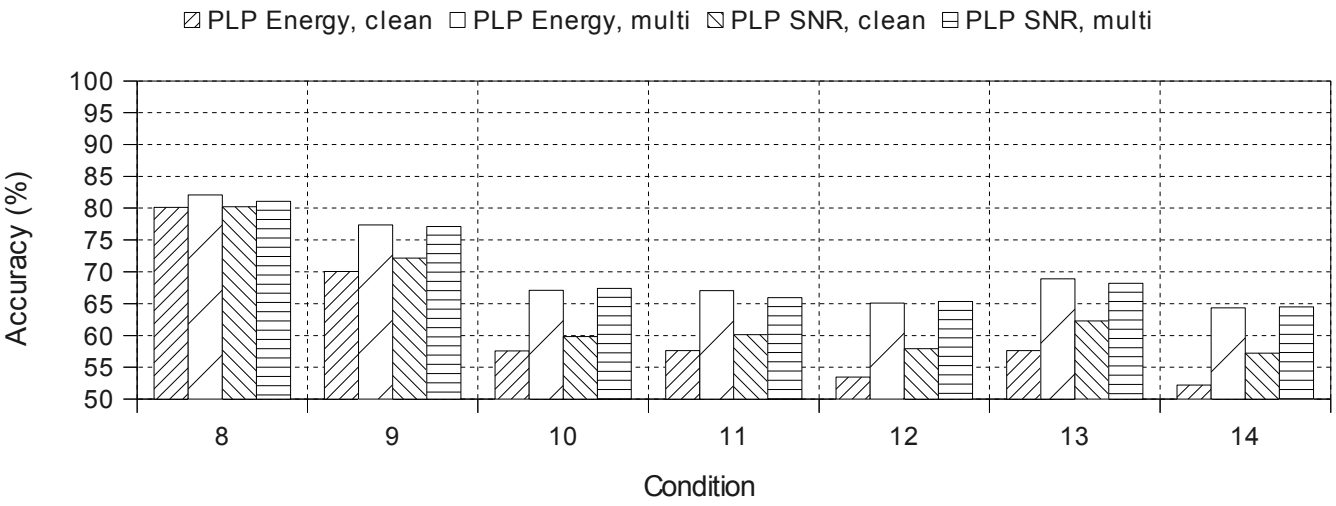

Figure 5: PLP results on aurora 4 database — second microphone.

6 Discussion

6.1 State of the art

The experiments show that SNR-spectrum based features can be beneficial in noisy environments when there is a mismatch between the training and testing conditions. Garner (2009) also showed that such features outperform various types of spectral subtraction. No other comparison is made with other noise robustness techniques. Rather, the use of standard databases means the results can be readily compared with those in the literature. No claim is made that the SNR-spectrum gives state of the art results. For instance, Li et al. (2007) report considerably better results on aurora 2.

6.2 Analysis

That the SNR-spectrum performs well is a curious result since there is not a large theoretical difference between SNR spectrum features and energy spectrum features when CMN is used. The difference is that the SNR spectrum features normalise before the filter-bank, whereas CMN works after it.

If the filter-bank weights for a single bin are denoted by $w_1, w_2, \ldots$, the SNR features presented to the decoder are of the form

$$\log\left(1 + w_1\xi_1 + w_2\xi_2 + \cdots\right), \tag{29}$$

whereas the energy based features are *closer* to the form

$$\log\left(1 + \frac{w_1(s_1 + n_1) + w_2(s_2 + n_2) + \cdots}{w_1 n_1 + w_2 n_2 + \cdots}\right). \tag{30}$$

In broadband noise, $\forall f : s_f \ll n_f$, both expressions clearly reduce to the same value ($\log 1$). However, if the noise is isolated to a particular bin, f, then only one term in the first expression will approach zero. In the second, the whole expression will reduce. It follows that the noises in the experimental conditions are suitably coloured for this effect to be significant.

These results are complementary to those of Lobdell et al. (2008), who also find advantages associated with AI features, albeit working after the filter-bank, and without cepstral normalisation.

6.3 PLP power law

One corollary of the aurora 2 experiments is that the cube root compression of Stevens (1957) normally used in PLP is not beneficial in the presence of noise. Whilst it is not the object of this study to investigate optimal PLP parameters, one hypothesis is as follows:

The compression affects the relative contribution of large and small spectral values in the LP calculation. Higher powers favour the higher values. The smaller power of 0.33 in PLP will enhance the contribution of smaller spectral values. The smaller values are likely to be noise. It follows that compression is in general not a noise robust operation. This issue is related to the SNR spectrum in that the SNR calculation can reduce noise peaks.

It can be tentatively concluded that additive noise is a more dominant concern than optimal compression in the present experimental conditions.

7 Conclusions

SNR-spectrum 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 combined with CMN and CVN perform well in noisy conditions, especially when the SNR is below 15dB.

The SNR-spectrum combined with the usual cepstral processing can be seen as an independent derivation of the articulation index. This also leads to insights into how to handle the noise tracker. Certainly the empirically optimal configuration is one with no hyperparameters. The SNR-spectrum is also closely related to features known to be beneficial in speech enhancement.

Experiments on artificial and restricted data give results that appear to generalise to real and less restricted data. Whilst no effort has been made to approach state of the art noise robustness figures, the SNR-spectrum appears complementary to techniques producing such results.

8 Acknowledgements

This work was supported by the Swiss National Science Foundation under the National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2). 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.

The author is extremely grateful to the four anonymous reviewers for their time and comments during the review process, especially regarding the presentation of results.

References

Alejandro Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, September 1990.

Alejandro Acero and Richard M. Stern. Environmental robustness in automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 849–852, April 1990.

Jont B. Allen. How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, October 1994.

Jont B. Allen. Consonant recognition and the articulation index. *Journal of the Acoustical Society of America*, 117 (4):2212–2223, April 2005.

Siu-Kei Au Yeung and Man-Hung Siu. Improved performance of aurora 4 using HTK and unsupervised MLLR adaptation. In *Proceedings of the International Conference on Spoken Language Processing*, Jeju, Korea, 2004.

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

Israel Cohen. Relaxed statistical model for speech enhancement and a priori SNR estimation. *IEEE Transactions on Speech and Audio Processing*, 13(5):870–881, September 2005.

Ángel de la Torre, Antonio M. Peinado, José C. Segura, José L. Pérez-Córdoba, Carmen Benítez, and Antonio J. Rubio. Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(3):355–366, May 2005.

Yariv Ephraim and David Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-32(6):1109–1121, December 1984.

ETSI. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. ETSI Standard 202 050, ETSI, 2002. V1.1.1.

Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29:254–272, April 1981.

Philip N. Garner. SNR features for automatic speech recognition. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, December 2009.

Philip N. Garner and John Dines. Tracter: A lightweight dataflow framework. In *Proceedings of Interspeech*, Makuhari, Japan, September 2010.

T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan. The AMI meeting transcription system: Progress and performance. In *Proceedings of the NIST RT06 Spring Workshop*, 2006.

Thomas Hain, Lukas Burget, John Dines, Philip N. Garner, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan. The AMIDA 2009 meeting transcription system. In *Proceedings of Interspeech*, Makuhari, Japan, September 2010.

Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

Hynek Hermansky and Nelson Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

Hans-Günter Hirsch and David 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"*, Paris, France, September 2000.

Guillaume Lathoud, Mathew Magimai-Doss, Bertrand Mesot, and Hervé Bourlard. Unsupervised spectral subtraction for noise-robust ASR. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, December 2005.

Guillaume Lathoud, Mathew Magimai-Doss, and Hervé Bourlard. Channel normalization for unsupervised spectral subtraction. IDIAP-RR 06-09, Idiap Research Institute, February 2006. URL http://publications.idiap.ch.

Jinyu Li, Li Deng, Dong Yu, Yifan Gong, and Alex Acero. High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, Kyoto, Japan, December 2007. IEEE.

Børge Lindberg. Danish SpeechDat-Car digits database for ETSI STQ aurora advanced DSR. Technical report, CPK, Aalborg University, January 2001. URL http://aurora.hsnr.de/download/sdc_danish_report.pdf.

Bryce E. Lobdell, Mark A. Hasegawa-Johnson, and Jont B. Allen. Human speech perception and feature extraction. In *Proceedings of Interspeech*, Brisbane, Australia, September 2008.

John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

Rainer Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9 (5):504–512, 2001.

Robert J. McAulay and Marilyn L. Malpass. Speech enhancement using a soft decision noise suppression filter. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(2):137–145, April 1980.

Pedro J. Moreno. *Speech Recognition in Noisy Environments*. PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, April 1996.

Pedro J. Moreno, Bhiksha Raj, and Richard 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*, volume 2, pages 733–736, Atlanta, US, May 1996.

Lorin Netsch. Description and baseline results for the subset of the Speechdat-Car German database used for ETSI STQ aurora WI008 advanced DSR frontend evaluation. STQ Aurora DSR Working Group input document AU/273/00, Texas Instruments, January 2001. URL http://aurora.hsnr.de/download/sdc_german_report.pdf.

N. Parihar, J. Picone, D. Pearce, and H. G. Hirsch. Performance analysis of the aurora large vocabulary baseline system. In *Proceedings of the 12th European Signal Processing Conference*, Vienna, Austria, September 2004.

Cyril Plapous, Claude Marro, Laurent Mauuary, and Pascal Scalart. A two-step noise reduction technique. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 289–292, Montreal, Canada, May 2004.

Christophe Ris and Stéphane Dupont. Assessing local noise level estimation methods: Application to noise robust ASR. *Speech Communication*, 34(1–2):141–158, April 2001.

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*, pages 225–228, 2002.

S. S. Stevens. On the psychophysical law. *Psychological Review*, 64(3):153–181, 1957.

Dirk Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3(2):151–167, April 1989.

Olli Viikki and Kari Laurila. Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization. In *Robust Speech Recognition for Unknown Communication Channels*, pages 107–110, Pont-à-Mousson, France, April 1997. ISCA.

Olli Viikki and Kari Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25:133–147, 1998.