IDIAP RESEARCH REPORT

COMBINING THE SNR SPECTRUM WITH A COCHLEAR MODEL

Philip N. Garner

Idiap-RR-14-2018

SEPTEMBER 2018

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

Combining the SNR Spectrum with a Cochlear Model

Philip N. Garner

April 5, 2016

Abstract

The SNR spectrum was previously introduced as a natural consequence of using cepstral normalisation in speech recognition; it is closely related to the articulation index of Fletcher. Motivated initially by a theoretical difficulty in frequency warping, the SNR spectrum is combined with a cochlear model, yielding a theoretically sound yet simple noise robust feature. Two time-domain cochlear model implementations, representing different recursive approximations, are compared. The combination of SNR spectrum, cochlear model, perceptual linear prediction and cepstral normalisation leads to an intuitive, efficient and effective feature that is mostly physiologically plausible.

Index Terms: cochlear model, SNR spectrum, gamma-tone filter.

1 Introduction

If automatic speech recognition is deployed in a noisy environment, it is well known that the implementation must explicitly take account of the background noise in order to prevent poor performance. Many techniques exist in the literature to enable this; the simplest is arguably the spectral subtraction of Boll [1]. The state of the art in noise robustness is probably in the body of literature derived from the vector Taylor series approach [2], or the deep learning approaches that are currently popular [3]. It is not the goal of this work to compete with such highly parametric approaches. Rather, we seek scientific insights into the way the human auditory system may achieve noise robustness, in the hope that this may in turn prescribe the "right" way to build components for such state of the art speech recognition systems.

The SNR spectrum [4, 5] was motivated by the fact that the robustness afforded by basic cepstral normalisation [6, 7, 8] is very difficult to improve upon in practice. Garner [4] showed that the use of cepstral mean normalisation (CMN) is equivalent to presenting a term of the form $\log(1 + \text{SNR})$ to the speech recognition decoder. Further, calculating this term from the outset rather than relying on CMN to produce it leads to theoretical and practical advantages. It was further demonstrated [5] that the SNR spectrum is particularly suited to linear prediction (in its perceptual form [9]), especially when combined with cepstral variance normalisation (CVN). The SNR cepstrum turns out to be very closely related (identical subject to linear transform) to the articulation index (AI) described by Allen [10, 11], who was also involved in its first use in a speech recognition context [12].

The SNR spectrum was defined in discrete Fourier transform (DFT) space. This was motivated in part because, under an additive Gaussian model, DFT outputs can be assumed to follow (complex) Gaussian distributions. This in turn makes the SNR derivation quite rigorous [4]. In practice, however, the SNR calculation is always followed by a frequency warp in the form of mel-spaced triangular filters. In principle, it would be better to calculate the SNR after the mel-filters as the samples would be better defined in a statistical sense; i.e., the larger bandwidth would lead to less variance. The distribution of a mel-warped variate can be calculated as a weighted sum of the distributions of the squared component DFT bins. However, whilst a squared DFT component is exponentially distributed, a sum of exponential distributions *with different parameters* is not easily tractable.

One solution to the tractability problem above is to implement the warped filter bank in the time domain; in the case of a linear filter, this would involve summing Gaussian variates, the result of which would remain Gaussian. This is in fact the approach taken by Lobdell et al. [12], but for different reasons. Many candidates exist for time domain filter forms, but an immediately apparent and compelling one is the gamma-tone filter bank which arises as a model of the cochlea [13].

In the remainder of the paper, the theory behind SNR features is reviewed motivating a time domain implementation. Two candidate gamma-tone implementations are described. They are tested with reference to the previous results of Garner [4, 5] with encouraging results.

An implementation is freely available via the tracter[1] and libssp[2] packages.

2 The SNR spectrum

Following the additive Gaussian model of [14], the value, t_f, of DFT bin f of a given frame can be assumed to be Gaussian distributed [5],

$$p(t_f \mid v_f) = \frac{1}{\pi v_f} \exp\left(-\frac{|t_f|^2}{v_f}\right). \tag{1}$$

In the case that only noise is present, the variance v_f is the noise variance ν_f. When both speech and noise are present, the variance is the sum $\sigma_f + \nu_f$, where σ_f is the speech variance. It was shown [4] that using the SNR spectrum rather than the absolute power spectrum (periodogram) has noise robustness properties. The maximum likelihood estimate of the SNR, ξ_f is

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

The value that is then input to the mel filter is then

$$1 + \hat{\xi}_f = \max\left(1, \frac{|t_f|^2}{\hat{\nu}_f}\right), \tag{3}$$

where $\hat{\nu}_f$ is the maximum likelihood value of the noise variance calculated in the intuitive way.

In order to rigorously calculate SNR in the mel domain, it is necessary to generate the distribution of a single mel bin. This amounts to applying a transformation of the form

$$m_f = \sum_{i=f-x}^{f+x} w_i |t_i|^2 \tag{4}$$

i.e., what is the distribution of a weighted sum of independent *but not identically* exponentially distributed variates? This is mathematically difficult.

A *heuristic* approach is to simply treat mel bins exactly as DFT bins are treated; this implicitly makes (reasonable) assumptions about the form of the distribution of mel bins. However, it is not rigorous.

3 Gamma-tone models

3.1 Background

The name gamma-tone was coined by Aertsen and Johannesma [15] as an approximation of a single sound element from a male grassfrog. More generally, it is a time domain description of a naturally occurring band-pass filter. One persuasive reason to favour it in auditory work is that it approximates reverse correlation measurements of the resonance of the basilar membrane; for a description with illustrations, see [16].

In the context of human cochlear models, the main text is by Patterson et al. [13], although a more detailed description is that of Holdsworth et al. [17]. The latter define the impulse response of a gamma-tone as

$$g_t(t) = t^{n-1} \exp(-2\pi b t) \cos(2\pi f_0 t + \phi). \tag{5}$$

where f_0 is the centre frequency and b is the bandwidth. They also point out that, for efficiency, the filter can be implemented (approximately) as a cascade of first order ($n = 1$) sections.

Gamma-tones have been used previously in speech recognition, notably by Schlüter et al. [18], who cite other work. In general, however, such cochlear models have not been shown to outperform the more common mel cepstra for general tasks (Schlüter et al. combine them with other features for good performance) and are not commonly used. Li and Huang [19] show that gamma-tone like features can benefit from further processing to achieve noise robustness for speaker identification.

In the context of the present work, the gamma-tone is a linear filter bank. Since each filter is linear, the distribution of filtered Gaussian variates remains Gaussian. This in turn means that essentially the same SNR calculations that are applicable to DFT bins can be applied to filter outputs, without further processing required for frequency warping.

[1] https://github.com/idiap/tracter
[2] https://github.com/idiap/libssp

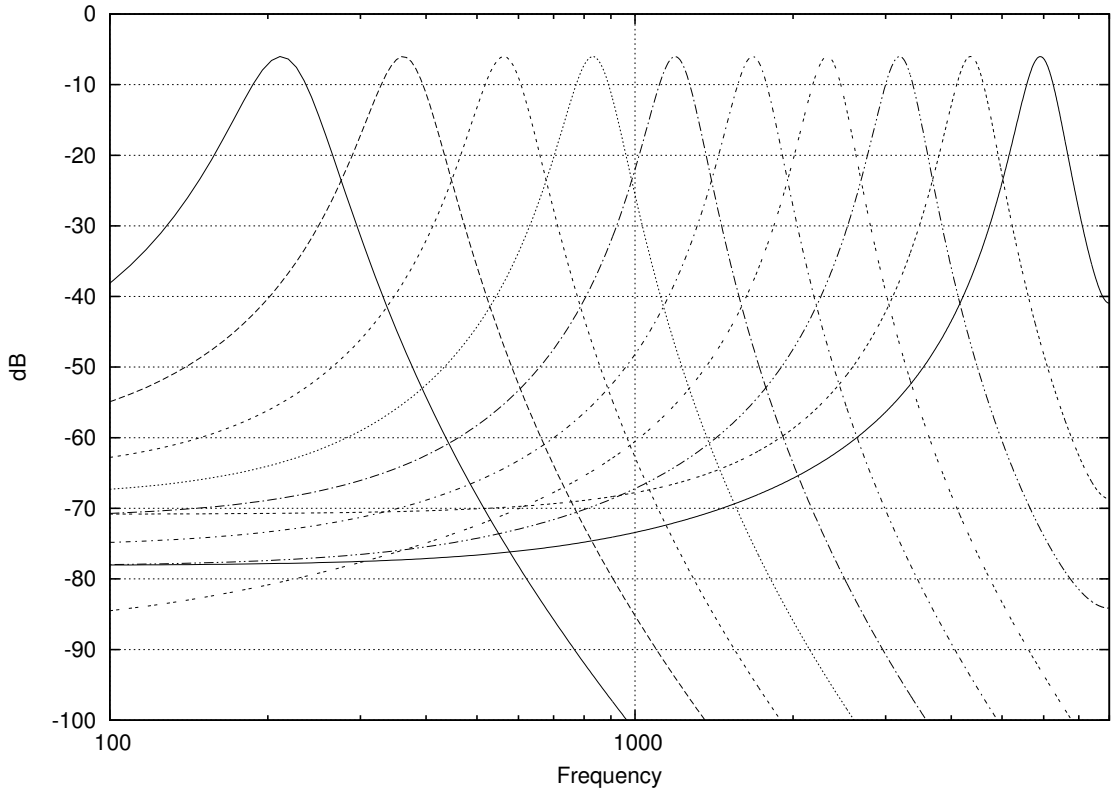

Figure 1: Example 10-channel frequency response of the cochlear filter bank of Holdsworth et al.

3.2 Holdsworth's implementation

Holdsworth et al. state that the filter can be implemented in three stages:

1. Frequency shift the array by $-f_0$ Hz
$$z_k = e^{-j2\pi f_0 kT} x_k. \tag{6}$$
where T is the sample period.

2. Pass through a first order recursive filter
$$w_k = w_{k-1} + \left(1 - e^{-2\pi bT}\right)(z_{k-1} - w_{k-1}). \tag{7}$$
This can be done multiple times for each order n.

3. Frequency shift the array by $+f_0$ Hz
$$y_k = \Re\left(e^{j2\pi f_0 kT} w_k\right). \tag{8}$$

Ma et al. [20] point out a trick to calculate the exponentials associated with the frequency shifts. Writing
$$e^{-j2\pi ft} = e^{-j2\pi f} e^{-j2\pi f(t-1)} \tag{9}$$
reduces each exponential to a multiplication for each k.

The parameters of the model are based on the equivalent rectangular bandwidth (ERB) scale. Following [21],
$$B_{ERB}(f) = 24.7(4.37 \times 10^{-3} f + 1), \tag{10}$$
The parameters can then be set [17] as:
$$b = \frac{1}{a_n} B_{ERB}(f) \tag{11}$$
$$a_n = \frac{\pi(2n-2)!\, 2^{-(2n-2)}}{(n-1)!^2}. \tag{12}$$

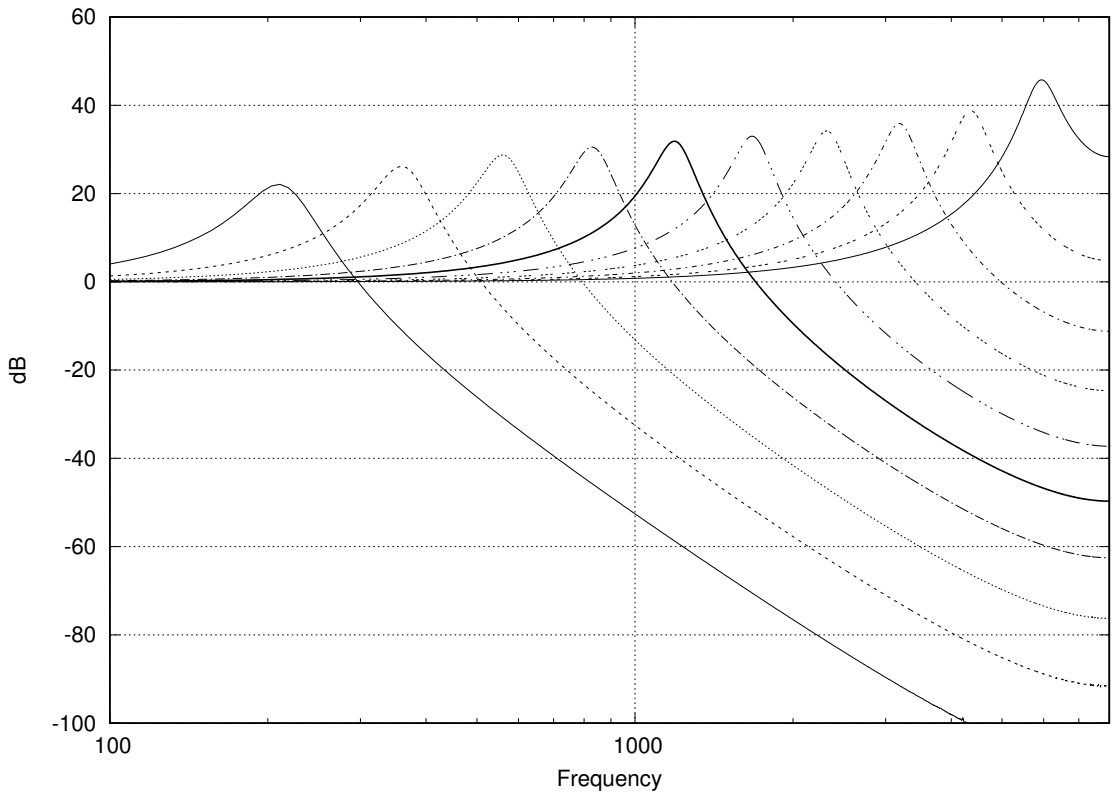

Figure 2: Example 10-channel frequency response of the APGF cochlear filter bank of Lyon

The centre frequencies are equally spaced on the "ERB rate" scale

$$\text{ERB rate} = 21.4 \log_{10}(4.37 \times 10^{-3} f + 1). \tag{13}$$

A filter bank implemented as described above is shown in figure 1.

3.3 Lyon's all-pole gamma-tone filter

Lyon [22, 23] points out that the gamma-tone form was originally described by Flanagan [24] to describe earlier measurements by von Békésy. Flanagan gives three approximations in the Laplace domain. His $F_3(s)$ is complicated, but has a gamma-tone time domain form; $F_1(s)$, however, although being complicated in the time domain, has a simple Laplace form:

$$F_1(s) = c_1 \beta^{4+r} \left(\frac{s+\epsilon}{s+\gamma} \right) \left[\frac{1}{(s+\alpha)^2 + \beta^2} \right]^2 e^{-sT}, \tag{14}$$

i.e., a high pass filter plus cascaded second order section (SOS). Lyon calls this the all-pole gamma-tone filter (APGF), and suggests that this is more appropriate as, at least in the right-hand quadrants of the z-domain, the asymmetry of the response is closer to that of the cochlea.

Lyon does not give an implementation, but one follows from the description above: To find a discrete version of this SOS, notice that (from tables)

$$\mathcal{L}\left(e^{-at}\sin(\omega t)\right) = \frac{\omega}{(s+a)^2 + \omega^2} \tag{15}$$

where $\mathcal{L}(\cdot)$ is the Laplace transform. Writing $t = kT$, the equivalent z-transform, $\mathcal{Z}(\cdot)$, is

$$\mathcal{Z}\left(e^{-akT}\sin(\omega kT)\right) = \frac{e^{-aT} z^{-1} \sin(\omega T)}{1 - 2e^{-aT} z^{-1} \cos(\omega T) + e^{-2aT} z^{-2}}. \tag{16}$$

Dividing the right hand side of equation 16 by its value at $z = 1$ (for unit gain at DC), and converting to a recursion formula,

$$y_k = (1 - 2e^{-aT}\cos(\omega T) + e^{-2aT})x_{k-1} + 2e^{-aT}\cos(\omega T)y_{k-1} - e^{-2aT}y_{k-2}. \quad (17)$$

Comparing equation 15 with Holdsworth's formulation of equation 5, ϕ can be arbitrarily set to cause cos to be sin, and

$$\omega = 2\pi f_0, \quad (18)$$

$$a = 2\pi b, \quad (19)$$

where b and f_0 can be set as above. Such a filter bank is illustrated in figure 2.

4 Experiments

4.1 General

Given the above description, a working hypothesis is that a performance gain may be achieved by replacing the periodogram and mel filterbank in a speech recognition front-end by a time-domain filter of the type described. In order to evaluate this hypothesis, experiments are presented involving incremental changes to a standard front-end.

4.2 Database

The aurora 2 task [25] 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 using different combinations of noise and channel. In previous work [5] it was shown that, although aurora 2 is artificial, it is a good indicator of performance on more realistic data.

4.3 Baseline performance

In an initial experiment, we assess the performance of the two gamma-tone filters compared to rather standard mel cepstra. The basic front-end begins with pre-emphasis by a filter with a single zero at $z = 1$. The signal is then framed into overlapping frames with a 10 ms period. The periodogram is then calculated for each frame, followed by a triangular mel-spaced filterbank. Finally, a logarithm and cosine transform yield cepstra of which the first 13 coefficients including "zeroth" term are retained. In the results, this case is referred to as "MFCC". Mean and variance normalisation are applied to the cepstra (CMVN) before augmenting with first and second order derivatives.

In the cochlear versions, the framing, periodogram and mel filter stages are replaced with time-domain filterbank followed by a framing and variance (energy) calculation per frame. The logarithm and cosine transform are then applied to the resulting vector, followed by CMVN and derivatives. In the results, the Holdsworth et al. implementation is referred to as "GF" and that of Lyon as "APGF". Note that the half-wave rectification typically associated with such models is not used; rather, the average frame energy is used as an estimate of the variance v_f.

A-priori, we may expect the cochlear versions to perform slightly worse than the mel baseline; this would be consistent with general feeling from the literature. We do not expect any extra noise robustness from uncompensated filters, although some authors do report it [26]. Neither do we expect any difference between cochlear implementations, except that the APGF should run more quickly.

Results are illustrated graphically in figure 3 for the basic front-end and figure 4, and summarised in table 1 for the three cases. In the figures, the first number in parentheses is the performance for clean test data, the second is the average of the 0 dB to 20 dB results. Since there is little difference between test sets, the tables summarise the latter metric averaged over test sets.

The opening hypothesis is borne out; the gamma-tone implementations do not perform as well as the standard front-end. In particular, although Lyon's APGF is competitive in matched conditions, it does not perform well in mismatched conditions. Further, (Holdsworth's implementation of) the conventional gamma-tone is not competitive.

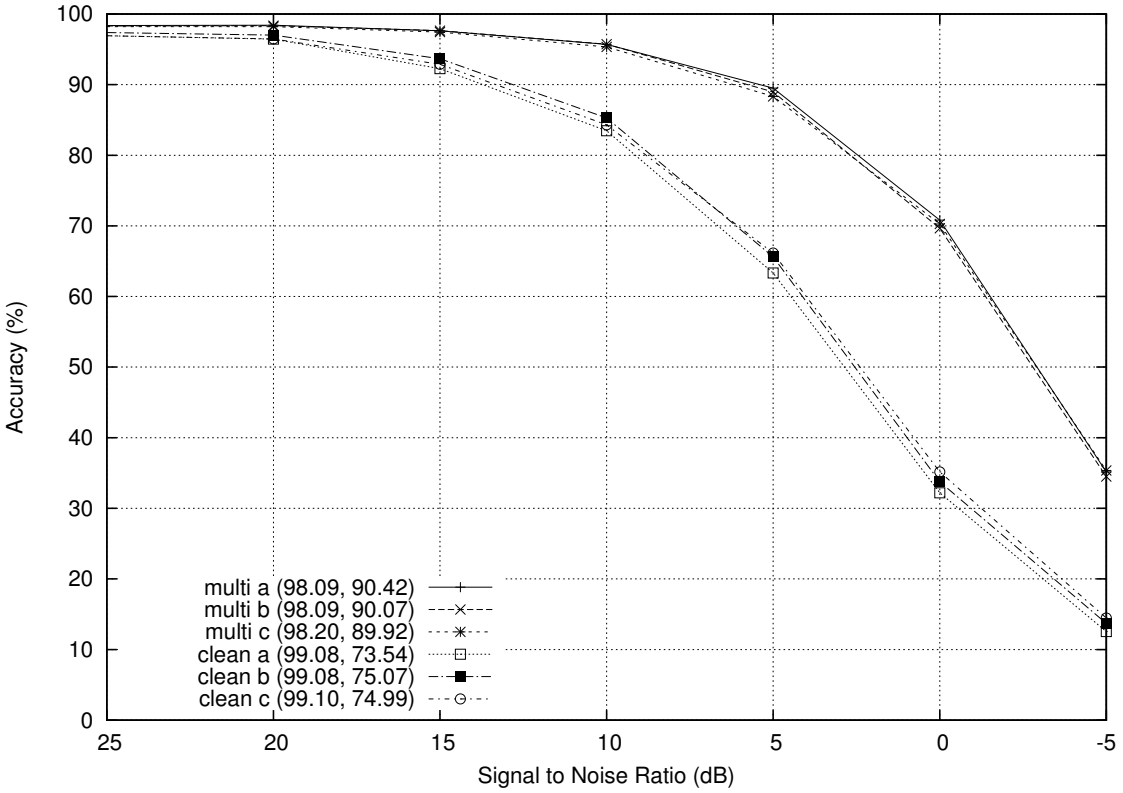

Figure 3: Baseline MFCC results.

Front-end	Multi	Clean
MFCC	90.1%	74.5%
GF	88.8%	57.7%
APGF	90.7%	70.4%

Table 1: Results for baseline tests.

4.4 SNR performance

In a second experiment we replace the inputs to the logarithm with SNR features following [4]. This case represents the main hypothesis of the paper, that the cochlear filter should perform better because of its adherence to the SNR estimation theory. An extra case is that of the heuristic "Mel" SNR, where the SNR is calculated on conventional mel filterbank outputs.

Front-end	Multi	Clean
MFCC	90.1%	82.0%
Mel	89.9%	80.6%
GF	89.9%	80.6%
APGF	90.5%	81.1%

Table 2: Results for SNR tests.

Results are summarised in table 2. Note that, whilst the results for matched conditions do not change much, there is a considerable improvement for the mismatched case. The performance gap between techniques is largely closed. It is difficult, however, to say that the hypothesis is demonstrated; whilst the APGF does outperform the heuristic mel approach, it is not by much.

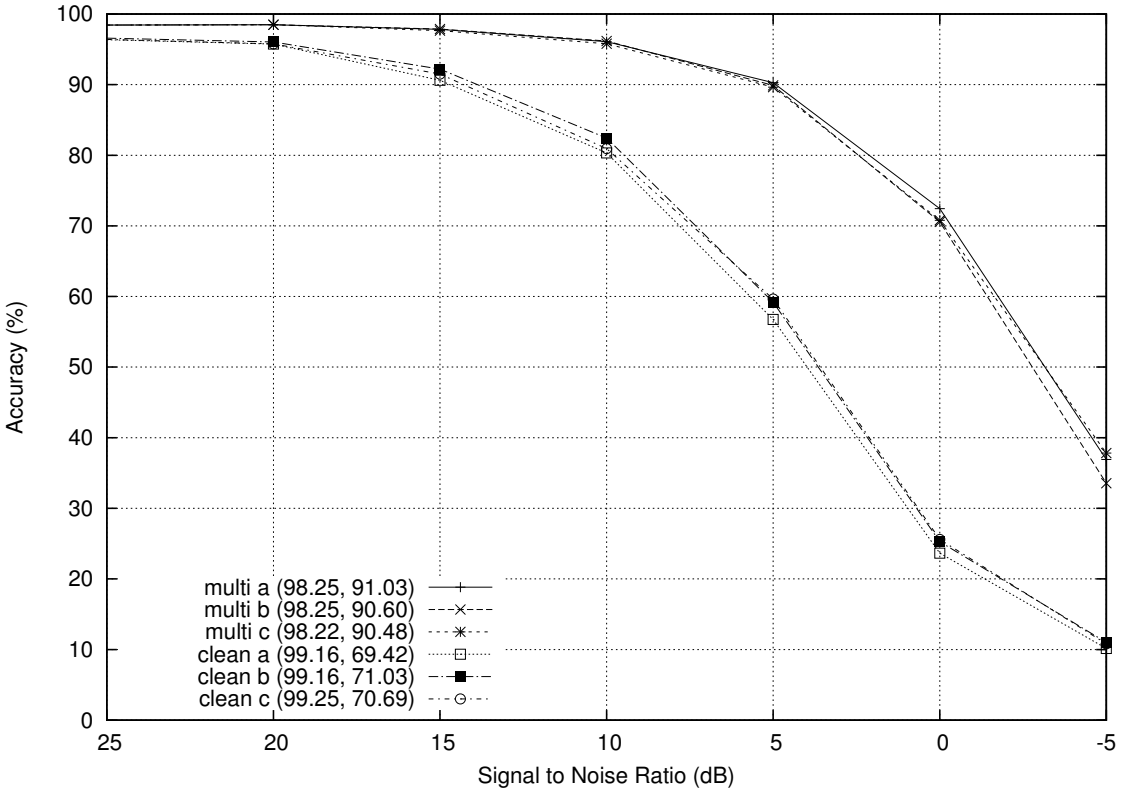

Figure 4: Baseline APGF results.

4.5 SNR-PLP performance

In a final experiment, we replace the cepstral calculation with a linear prediction stage as in [5]; i.e., a cosine transform before the logarithm yields the (warped) autocorrelation from which LP cepstra follow. In previous work [5], it was shown that this led to a significant performance improvement on clean training conditions. We would expect a-priori to see such an improvement in the cochlear filter case too.

Front-end	Multi	Clean
MFCC	90.5%	84.5%
Mel	91.0%	83.9%
GF	90.9%	84.5%
APGF	90.8%	84.9%

Table 3: Results for PLP-SNR tests.

Results are summarised in table 3. This is particularly interesting in that, whilst all techniques show an improvement, both gamma-tone solutions perform equally or perhaps better than the conventional cases. The clean result lends weight to the opening hypothesis, although again the results are probably too close.

4.6 Hyper-parameters

Although in general we try not to optimise hyper-parameters, some issues arise. The noise estimation is based on the minima tracking approach of [27, 28]. This in turn requires a multiplicative correction C; in the DFT case, $C \approx 11$. The AI, however, also has a correction factor that multiplies the SNR. Without an analysis of the noise estimator, it is impossible to distinguish the two factors. It was reported by [5] that setting $C = 1$ took care of both factors; it was hypothesised that they are related. In the present study, whilst $C = 1$ was used in the MFCC case, the heuristic mel and (AP)GT cases required $C = 0.5$; this was found heuristically. An analysis of this factor is a matter for future work.

The APGF implementation uses the bandwidth calculations appropriate for the GF case with two SOSs

and n set to 2. The GF case has order $n = 4$. Whether the the GF bandwidth is appropriate for the APGF case is also a matter for future work.

5 Conclusion

The opening hypothesis of this study was that a time domain warped filter-bank could work better than a mel-binned one because the SNR calculation could be more rigorous. In fact, although there is some evidence that this has been demonstrated, the evidence is not large. Further, there are other small differences between implementations such as mel vs. ERB warping and filter shapes that could account for the differences.

By contrast, however, there was a small expectation that time-domain filters may not perform as well as conventional MFCCs. Whilst this is true for the uncompensated case, it is not so when noise compensation is introduced. In particular, the APGF is able to even outperform the conventional case. In fact, although it was not an aim of the study, figures 3 and 4 show that the APGF can also outperform the conventional case for clean test data.

A tentative conclusion follows that the combination of APGF, SNR spectrum (AI), linear prediction and cepstral normalisation is a promising component in a state of the art speech recognition system. It is also appealing not only in that it is mostly physiologically plausible, but also because it has a sound mathematical basis.

6 Acknowledgements

This work was supported in part by the H2020 SUMMA project, grant agreement 688139/H2020-ICT-2015, and by the Swiss NSF SIWIS project, grant CRSII2 141903.

References

[1] 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.

[2] J. Li, L. Deng, D. Yu, Y. Gong, and A. 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: IEEE, December 2007.

[3] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 2013, pp. 7398–7402.

[4] P. N. Garner, "SNR features for automatic speech recognition," in *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, December 2009.

[5] ——, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Communication*, vol. 53, no. 8, pp. 991–1001, October 2011.

[6] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 254–272, April 1981.

[7] 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*. Pont-à-Mousson, France: ISCA, April 1997, pp. 107–110.

[8] ——, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[9] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[10] J. B. Allen, "How do humans process and recognize speech?" *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, October 1994.

[11] ——, "Consonant recognition and the articulation index," *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2212–2223, April 2005.

[12] B. E. Lobdell, M. A. Hasegawa-Johnson, and J. B. Allen, "Human speech perception and feature extraction," in *Proceedings of Interspeech*, Brisbane, Australia, September 2008.

[13] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception, Proceedings 9th International Symposium on Hearing*, Y. Cazals, L. Demany, and K. Horner, Eds., 1992, pp. 429–446.

[14] 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.

[15] A. M. H. J. Aertsen and P. I. M. Johannesma, "Spectro-temporal receptive fields of auditory neurons in the grassfrog: I. characterization of tonal and natural stimuli," *Biological Cybernetics*, vol. 38, no. 4, pp. 223–234, November 1980.

[16] E. de Boer and H. R. de Jongh, "On cochlear encoding: Potentialities and limitations of the reverse-correlation technique," *Journal of the Acoustical Society of America*, vol. 63, no. 1, pp. 115–135, January 1978.

[17] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "SVOS final report. (part A: The auditory filter bank)," February 1988, annex C: Implementing a GammaTone Filter Bank. [Online]. Available: http://w3.pdn.cam.ac.uk/groups/cnbh/research/publications/pdfs/SVOSAnnexC1988.pdf

[18] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HI, USA, April 2007, pp. 649–652.

[19] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1791–1801, August 2011.

[20] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Communication*, vol. 49, no. 12, pp. 874–891, December 2007.

[21] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1–2, pp. 103–138, August 1990.

[22] R. F. Lyon, "The all-pole gammatone filter and auditory models," Apple Computer, Inc., Cupertino, CA 95014 USA, Tech. Rep., 1996, draft. [Online]. Available: http://www.dicklyon.com/tech/Hearing/APGF_Lyon_1996.pdf

[23] ——, "All-pole models of auditory filtering," in *Proceedings of the International Symposium on Diversity in Auditory Mechanics*, E. R. L. et al., Ed., 1997, pp. 205–211, university of California, Berkeley 24-28 June 1996. [Online]. Available: http://www.dicklyon.com/tech/Hearing/DAMReprint-Lyon.pdf

[24] J. L. Flanagan, "Models for approximating basilar membrane displacement," *The Bell System Technical Journal*, vol. 39, no. 5, pp. 1164–1191, September 1960.

[25] 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"*, Paris, France, September 2000.

[26] A. Adiga, M. Magimai.-Doss, and C. S. Seelamantula, "Gammatone wavelet cepstral coefficients for robust speech recognition," in *Proceedings of IEEE TENCON*, Sydney, Australia, October 2013.

[27] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Communication*, vol. 34, no. 1–2, pp. 141–158, April 2001.

[28] 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