ANALYTICAL AND ITERATIVE APPROACHES TO THE EQUALISATION OF SUB-BAND ERRORS IN SPEECH AND SPEAKER RECOGNITION

Roland Auckenthaler† *and John S. Mason*‡
†Department of Electronics,Technical University Graz,
Inffeldgasse 12,A-8010 GRAZ, AUSTRIA
‡Department of Electrical & Electronic Engineering,
University of Wales Swansea, SA2 8PP,UK
email: {eeaucken, J.S.D.Mason}@swansea.ac.uk

ABSTRACT

Recent work in both speech and speaker recognition has shown some interesting apparent benefits of sub-band processing: dividing the acoustic band into sub-units to give multiple stream (sub)-classifiers.

In this paper we extend our recent work on sub-band error equalisation by considering 4 separate cases from combinations of male and female speaker sets in the context of speech and speaker recognition.

We show that sub-band error equalisation can be achieved by changing the conventional mel frequency warping function. This can also reduce the overall error rates significantly: in the case of the female speaker set the rate is reduced by over 50%.

1 INTRODUCTION

Recent work in speech [1] [2] and speaker recognition [3] [6] has shown benefits in dividing the audio band into sub-bands and performing classification on each sub-band.

Potential benefits of this sub-band approach include robustness against narrow-band noise, closer simulation of human perception [4], and the possibility of tailoring the processing in time and frequency. It is also possible that intra-speaker variation includes cross-band drift resulting in damaging miss-alignments. An apparent disadvantage would seem to be the loss of inter-band information. The most direct approach to assess the net effect of such postulations is via experimentation. To this end Besacier [3] [5] has shown distinct benefits of pruning in time and frequency. Furthermore, in our previous work [6] [7], in attempting to equalise sub-band errors, we have demonstrated that the use of two sub-bands consistently equals or out-performs single band operations, without any time or frequency pruning.

In this paper we extend our previous work on sub-band error equalisation to include both speech and speaker recognition, with the important bi-product of comparing combined sub-band classification error rates with the conventional single band case.

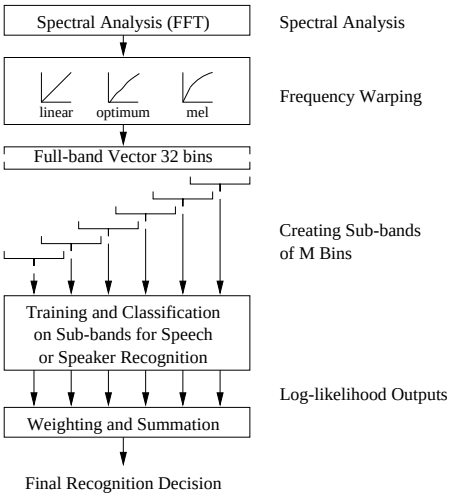


Figure 1: Concept of a sub-band recognition system

2 SUB-BAND EQUALISATION

Our hypothesis is that an arrangement of sub-bands that results in equal levels of discrimination would be beneficial; for example when a given sub-band is to be omitted because of temporary signal degradation then the remaining sub-systems could remain equally weighted on recombination. And if no prior knowledge of noise degradation can be assumed, then equalising bands would seem an optimum choice.

In previous work [6], we have approached equalisation by relating sub-band errors to a frequency warping or scaling function. In speech recognition the very popular mel scale is a standard for such warping. Here we adopt similar mel-like functions to meet the goal of sub-band error equalisation in the context of both speech and speaker recognition, and for male and female speaker sets. Ideally the equalisation should be achieved while maintaining (or improving) recognition performance on the overall conventional full-band system.

Figure 1 illustrates the arrangement. The frequency warping function is applied to the output of the FFT followed by a smoothing function which combines adjacent frequency bins, reducing the number in this case from

128 to 32. This is shown to be a good choice for speaker recognition by [8]. Each original time frame contains 256 samples spanning 32ms.

3 INDIVIDUAL SUB-BAND ERROR RATES

Initially two cases are considered: a linear function so that each of the resultant sub-bands contains just M of the 32 bins, and the popular mel case [9] where the warping function takes on a log form at frequencies above 1kHz, with much wider bins at high frequencies. The optimum warping function in the current situation is defined as one that gives equal (and low) error rates as a sliding window of M bands moves across the 32 bands.

Closed-set experiments are conducted in the context of both *speech* and *speaker* recognition, using the BT Miller database and single-word token, isolated digits. Two speaker sub-sets are considered: one of 20 male speakers and one of 14 female speakers. Ten versions recorded over the first 2 sessions are used for training and 15 versions recorded over 3 subsequent sessions are used for testing. In both cases of speech and speaker recognition a single dynamic-time-warped (DTW) model is used to represent each class, and in each case just one model per class is used ie speaker-independent speech recognition.

3.1 Assessment of Errors

Results for the mel and linear scales are shown in Figure 2 for M=5. The plots on the left-hand side are sub-band error rates (speaker identification or speech recognition). Comparison should be made with care since centre frequencies and bandwidths differ in the two cases. The mel scale gives narrower low-frequency sub-bands, thus their profiles are all shifted to the left slightly, ie they begin and end at a lower centre frequencies.

The plots on the right-hand side of the page are designed to illustrate the departure from equal-error sub-bands. Consider:

$$E_n = \sum_{i=1}^{n} e_i \quad and \quad \bar{E_n} = \frac{E_n}{E_N} \tag{1}$$

where e_i is the sub-band error, N the total number of sub-bands and E_n the sum of e_1 to e_n. Then

$$\Delta\bar{E_n} = \bar{E_n} - \frac{n}{N} \tag{2}$$

represents the departure of the normed error $\bar{E_n}$ to an optimal straight line. For the desired warping function with equal contributions of e_i these profiles would coincide with the horizontal axis. Thus, in this respect it can be seen that in two of the four cases the mel and linear scale are opposites: top right, *male speaker* recognition, and bottom right, *female speech* recognition.

The top 4 figures relate to the male speaker set, while the bottom 4 relate to the female set. On a general note, in three out of the four cases it can be seen that the mel-scale gives a peak in the error profiles in the region leading up to 1000Hz, followed by lower error rates at higher frequencies. This suggests that the mel scale might be too narrow in this region.

The profiles in the top pair of figures (male speaker recognition) are very similar to our previous findings, [6] where we used a VQ classifier, rather than the DTW approach adopted here. The second pair of plots - in b) - suggest that for *male speech* recognition, the mel function is close to optimum, though the error rate is relatively high at low frequencies with the lowest rate at about 1500Hz.

The bottom 4 figures under c) and d) are for the female set, and show some interesting contrasts to those for the male set. Beginning with speaker recognition,the profiles for the mel scale do in fact show similar trends to those for the male set, with a peak in the error profiles between 50Hz and 1000Hz. This general similarity is corroborated in the bell curves on the right. However, a potentially important difference occurs in the linear case, where for the female set in speaker recognition, this seems close to the optimum. Finally, the last pair of figures, those under d), are for speech recognition for the female set, and here it is clear that the picture is similar to that for the male set for speaker recognition, with mel and linear being some kind of opposites, with the optimum somewhere between the two.

In terms of flat sub-band error profiles, reasonable results are obtained with mel scale and male set speech recognition and linear scale female speaker recognition. This leaves the other two combinations as candidates for a new warping function somewhere between mel and linear, as illustrated by the top and bottom plots in Figure 2.

3.2 Sub-band Errors after Equalisation

Strategies we have examined for equalisation include an analytic approach [6] where the error rate is assumed proportional to the inverse of the band-width. Then, for the *male speaker* recognition task, replacing the standard mel function of the form:

$$f_{mel} = \begin{cases} f & : \quad f < 1000Hz \\ 2595\log(1+\frac{f}{700}) & : \quad f \geq 1000Hz \end{cases} \tag{3}$$

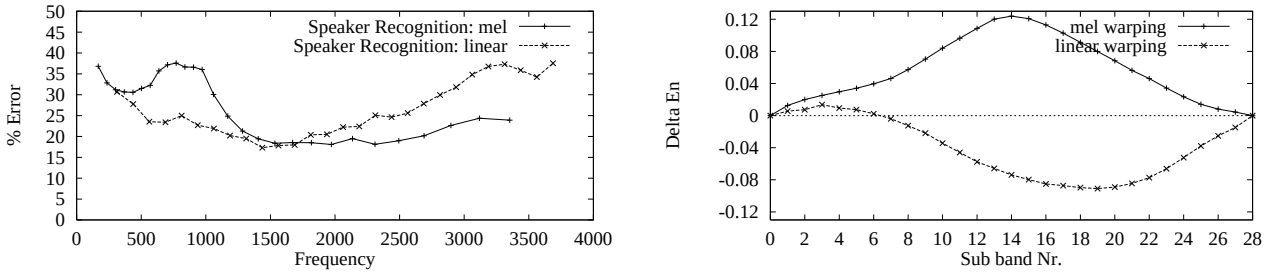
with a mel-like function:

$$f_{ml} = \begin{cases} \frac{4}{3}f & : \quad f < 1500Hz \\ 4912\log(f+100) - 13738 & : \quad f \geq 1500Hz \end{cases} \tag{4}$$

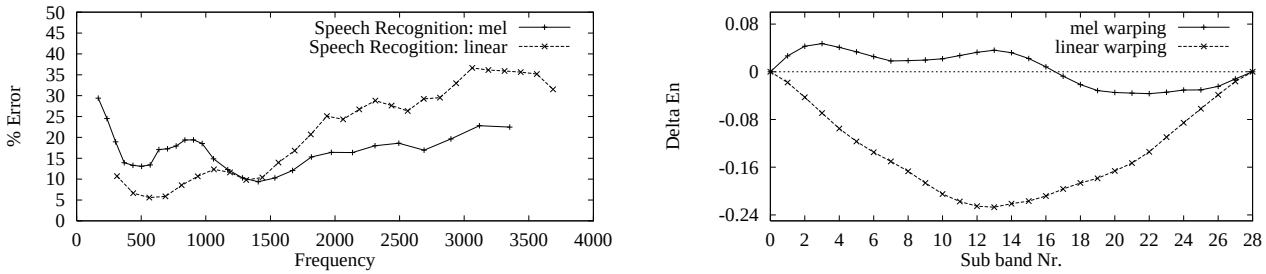
gives improved performance.

An alternative is to optimise directly on an error function cost using an iterative approach [7]. This leads to the piece-wise linear function:

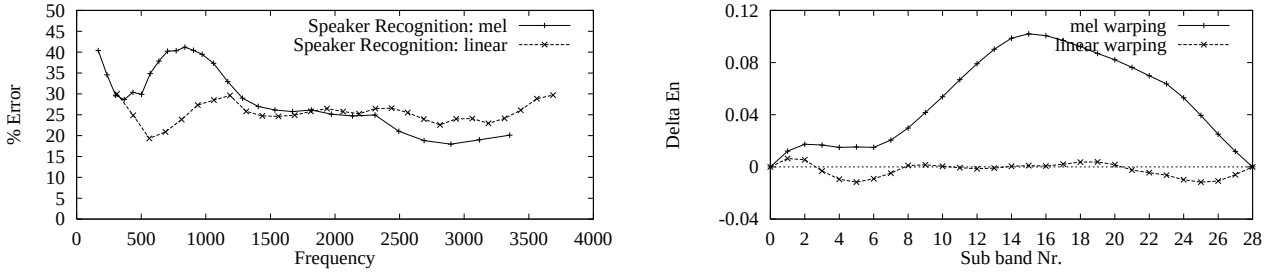
$$f_{eq_m} = \begin{cases} f & : \quad f \leq 1125Hz \\ \frac{4}{3}f - 375 & : 1125 < f \leq 2159Hz \\ f + 345 & : 2159 < f \leq 2659Hz \\ \frac{4}{5}f + 877 & : 2659 < f \leq 2972Hz \\ \frac{8}{11}f + 1092 & : 2972 < f \leq 4000Hz \end{cases} \tag{5}$$



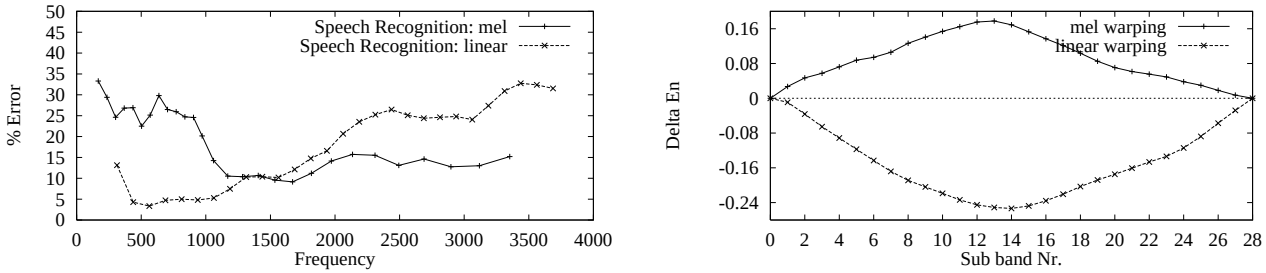
a) *Speaker* Recognition, Male Speaker Set



b) *Speech* Recognition, Male Speaker Set



c) *Speaker* Recognition, Female Speaker Set



d) *Speech* Recognition, Female Speaker Set

Figure 2: Sub-band error profiles (left) and $\Delta\bar{E}_n$ (right) for mel and linear warping for *speech* and *speaker* recognition

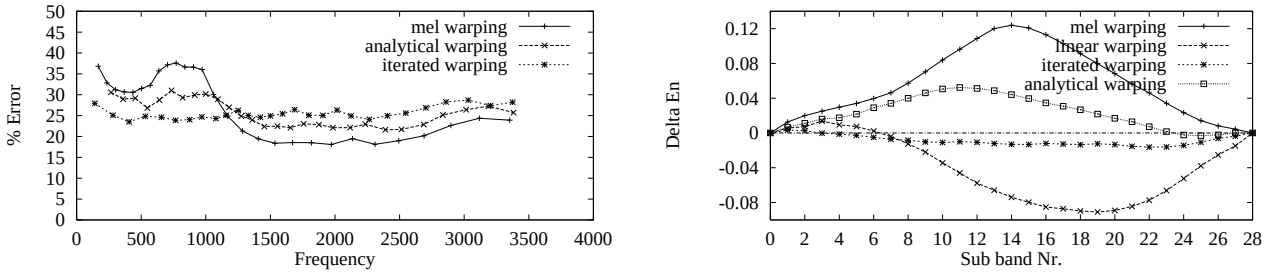
These experiments are repeated here using DTW rather than VQ, and the results are presented in Figure 3.

The iterative solution is seen to give a much flatter profile and thus is the approach adopted here for equalisation of the second combination, namely a *female speech* recognition context, the resultant function for which is:

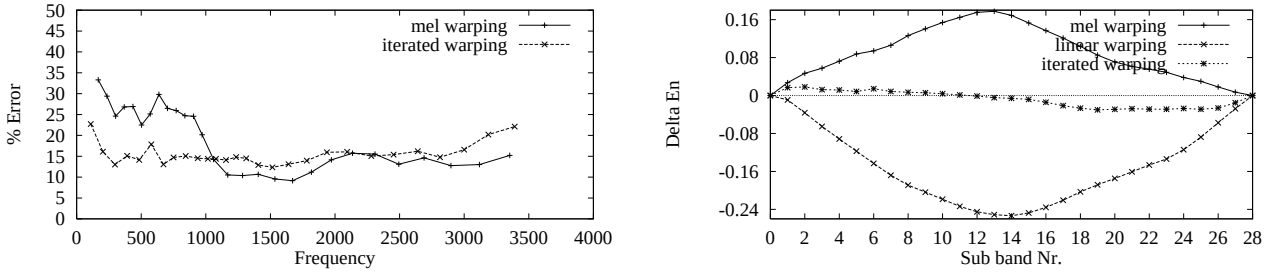
$$f_{eq_f} = \begin{cases} \frac{4}{3}f & : \quad f \leq 1313Hz \\ 2f - 875 & : \quad 1313 < f \leq 1563Hz \\ f + 688 & : \quad 1563 < f \leq 1938Hz \\ \frac{16}{23}f + 1278 & : \quad 1938 < f \leq 3375Hz \\ \frac{3}{5}f + 1600 & : \quad 3375 < f \leq 4000Hz \end{cases} \quad (6)$$

The sub-band error profiles are shown in 3b).

Finally we consider the overall performance of systems with a small number of *non-overlapping* sub-bands, with score combination prior to the decision stage. Results are shown in Table 1. While there are only small differences in the results for the male speaker set for both speech and speaker recognition, this is not the case for the female speaker sets where significant improvements over standard mel are obtained using either a linear scale for speaker recognition, or an equalised scale, Equation 6, for speech recognition.



a) *Speaker* Recognition, Male Speaker Set



b) *Speech* Recognition, Female Speaker Set

Figure 3: Comparison of sub-band error profiles (left) and $\Delta\bar{E}_n$ (right) for mel, linear and iterated warpings

Task	Set	Warp.	1Band	2Bands	4Bands
Speaker Recogn.	male	f_{mel}	2.90	2.83	2.33
	male	f_{eq_m}	2.90	2.27	2.53
	female	f_{mel}	9.81	8.29	8.43
	female	f_{lin}	4.43	4.33	4.10
Speech Recogn.	male	f_{mel}	0.87	1.07	1.10
	male	f_{eq_m}	1.07	1.03	1.43
	female	f_{mel}	0.29	0.19	0.57
	female	f_{eq_f}	0.19	0.14	0.24

Table 1: Recognition errors for different speaker sets in speech and speaker recognition for 1,2 and 4 sub-bands

4 CONCLUSIONS

The goal of this paper was to examine sub-band error rates in the context of speech and speaker recognition, and to attempt to equalise these rates across the bands. We have shown that frequency warping functions such as the standard mel can be changed to give the desired equalisation without degradation in overall performance. In fact using a frequency scale which gives the latest sub-band error profile in the case of the female speaker set gives significant improvement in performance, error rates falling from 9.8% to 4.4% and 0.29% to 0.19% for speaker and speech recognition respectively.

References

[1] H. Hermansky, S. Tibrewala, and M. Pavel. Towards ASR on partially corrupted speech. In *Proc. ICSLP*, volume 1, pages 462–465, 1996.

[2] H. Bourlard and S. Dupont. A new ASR approach on independent processing and recombination of partial frequency bands. In *Proc. ICSLP*, volume 1, pages 426–429, 1996.

[3] L. Besacier and J. Bonastre. Subband approach for automatic speaker recognition. In *Proc. AVBPA*, pages 195–202, 1997.

[4] J. B. Allen. How do humans recognize speech? *IEEE Trans. on ASSP*, 2(4):pages 567–577, October 1994.

[5] L. Besacier and J. F. Bonastre. Time and Frequency Pruning for Speaker Identification. *RLA2C*, page 106, 1997.

[6] R. Auckenthaler and J.S. Mason. Equalising Sub-Band Error Rates in Speaker Recognition. In *Proc. EUROSPEECH*, volume 5, pages 2203–2206, 1997.

[7] R. Auckenthaler and J. S. Mason. Warping Function for Sub-band Error Equalisation in Speaker Recognition. *RLA2C*, page 194, 1997.

[8] L. Xu and J.S. Mason. Optimization of perceptually-based spectral transforms in speaker identification. In *Proc. EUROSPEECH*, pages 439–442, 1991.

[9] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on ASSP*, pages 357–366, 1980.