EQUALIZING SUB-BAND ERROR RATES IN 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 ASR shows that band splitting, forming multiple paths with recombination at the decision stage, can give recognition accuracy comparable with the conventional full-band approach. One of the many interesting questions with band-splitting relates to the bandwidths of each sub-band, and the use of frequency warping functions such as mel. This paper examines the use of mel and linear frequency scales in the context of band-splitting and speaker recognition. We demonstrate how sub-band error profiles can lead to a new scale, which is between linear and mel, giving both an equalised sub-band error profile and an improved overall recognition accuracy.

1. INTRODUCTION

This paper is concerned with splitting the conventional acoustic representation into sub-band units and processing these separately, with recombination at the decision stage. This idea has been investigated recently in the context of speech recognition [1] [2] and the complementary one of speaker recognition [3].

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. Here, we focus on the band splitting itself and filter-bank analysis to give equalised error profiles across sub-bands.

In order to attain the full potential of the sub-band recognition approach, it might well be necessary to distribute the sub-band errors appropriately. For example in the case of noise robustness, it would be advantageous to remove (or de-weight), those bands subjected to high-levels of noise. Ideally it would be desirable to eliminate only those bands which contain no discriminating information, retaining the remainder, suitably weighted, according to usefulness.

Here our hypothesis is that it would be beneficial to arrange for sub-bands to have approximately equal levels of discrimination. Such a balance might well lead to an optimum arrangement of sub-band units. This balance across the bands is to be achieved by a suitable warping function applied in the same manner as the popular mel scale. In fact, our initial experiments begin with the two cases of mel and linear scales.

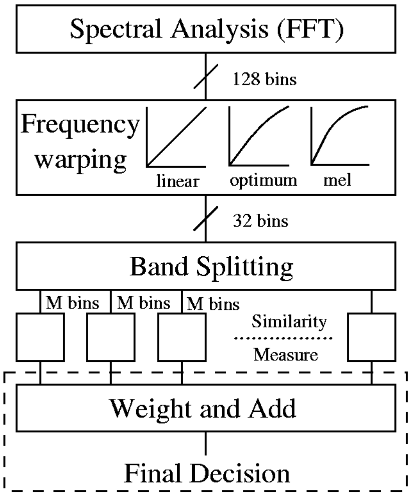

Figure 1: Concept of a sub-band recognition system

Figure 1 shows the band splitting approach. Conventional filter-bank analysis, including a change of frequency resolution, follows the FFT, resulting in a reduced number of spectral bins to represent the speech in the spectral domain. Here we use throughout 32 bins, since this number is shown to be a good choice for speaker recognition [5].

For sub-band operation, these 32 bins are divided into units each of M bins, leading ultimately to multiple threads for classifying, with combination before the decision stage.

However, initially recognition experiments are performed using individual sub-band units across the frequency range, analogous to a moving average, with a maximum frame-rate. Results for mel and linear scales are shown in Figure 2 for M=5. Comparison should be made with care since centre frequencies and bandwidths differ in the two cases. The recognition experiments are based on a closed-set, 20 speaker, digit-dependent, single-word token, speaker identification with 10 versions training and 15 different versions for testing, derived from the BT Millar database.

It can be seen (Figure 2), that the linear scale gives an almost monotonic rise in error rate with frequency, while both linear and mel scale profiles exhibit a peak between 700 and 900 Hz. This peak for the mel case is much larger because, although the mel scale is itself linear in this region, the bandwidth of the mel filters is much lower than for the corresponding linear case, to compensate for the much broader mel bands at higher frequencies.

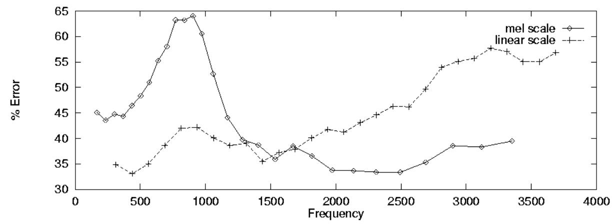

Figure 2: Recognition errors for M=5-bin sub-bands vs. frequency

Our goal is to flatten the profiles in Figure 2, so that the level of discrimination per unit frequency bin is constant while maintaining, or even improving, the overall recognition performance.

2. A LEVEL SUB-BAND ERROR RATE

Consider the two profiles in Figure 2. Each exhibits some almost linear, constant slope sections which suggest that to a first approximation error rates can be regarded as inversely proportional to the bandwidth of of sub-bands, ie:

$$e_i \propto \frac{1}{Bandwidth_i} \tag{1}$$

where e_i is the experimental evaluated recognition error for sub-band i. Below, this assumed relationship is used in attempting to derive an equal error profile.

First though, we contrast further the mel and linear error profiles. The regions above and below about 1500 Hz suggest that a compromise between the mel and the linear scale might well lead to flatter profiles. To illustrate this we integrate and normalise the error profiles.

Consider:

$$E_n = \sum_{i=1}^{n} e_i \quad and \quad \bar{E_n} = \frac{E_n}{E_N} \tag{2}$$

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

$$\Delta\bar{E_n} = \bar{E_n} - \frac{n}{N} \tag{3}$$

represents the departure of the normed error $\bar{E_n}$ to an optimal straight line. This departure is shown in Figure 2 for linear (lower profile) and mel (upper profile) scaled bins.

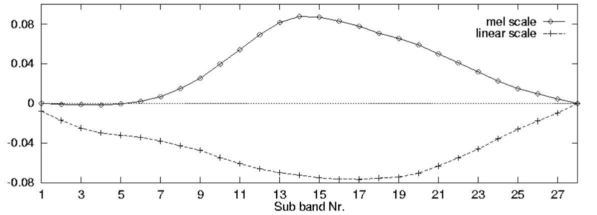

Figure 3: Normalised departure measure, $\Delta\bar{E_n}$, for mel (upper) and linear (lower) filter-banks

For the desired warping function with equal contributions of e_i the profiles in Figure 3 would coincide with the horizontal axis. Thus, in this respect it can be seen that the mel and linear scale are opposites.

We investigate these observations and address the question of whether such alternative scales can be found which both flatten the error profiles and lead to competitive recognition performance.

3. A NEW WARPING FUNCTION

Consider a scaler function, $T_{mel}(f)$, which links the two error profiles in Figure 2 such that

$$e_{mel}(f) = T_{mel}(f)e_{lin}(f) \tag{4}$$

where $T_{mel}(f)$ is simply the ratio of the profiles at each frequency. Clearly, such a function is directly linked to both the error profiles (as is obvious from equation (4)) and also the corresponding warping function, in this case the standard mel warping function, f_{mel}:

$$f_{mel} = \left\{ \begin{array}{ccc} f & : & f < 1000Hz \\ 2595\log\left(1 + \frac{f}{700}\right) & : & f \geq 1000Hz \end{array} \right. \tag{5}$$

We propose using such a link to establish a warping function, similar in form to the mel warping, but which results in a flat equalised error profile in place in $e_{mel}(f)$. We again make use of the linear error profile and the associated scalar ratio function, $T_{equal}(f)$, which relates $e_{lin}(f)$ to our desired flat profile:

$$e_{equal}(f) = T_{equal}(f)e_{lin}(f) \tag{6}$$

The task now is to determine the associated warping function, f_{equal}, which leads to $e_{equal}(f)$ in an analogous manner that the mel warping function, f_{mel}, leads to $e_{mel}(f)$.

The hypothesis is that

$$T_w(f) \propto \frac{1}{Bandwidth(f)} \propto \frac{df_w}{df} \tag{7}$$

where the scaler ratio function $T_w(f)$ relates to a given warping, f_w. This hypothesis comes from the observation that, in the case of the standard mel warping function, the inverse of the bandwidth is approximately proportional to the slope of the warping function. The case here is that this is a more general relationship, not applicable just to mel warping.

The error profile is evaluated here for sub-bands with M=5 bins. As M increases the profile becomes smoother, and as M is reduced the results become less informative. If equation (7) holds, then it follows that the warping function, f_{equal}, can be derived from integrating T_{equal} i.e. the inverse of $e_{lin}(f)$.

However this is not simple since it is dependent upon M and, taking the case M=5 for example, does not generally have an easily integrate-able form.

Therefore we test the hypothesis expressed in equation (7) in the case of the standard mel, f_{mel} (equation (5)), and also a second simple two linear case warping function, f_{2lin}:

$$f_{2lin} = \left\{ \begin{array}{ccc} 1.2f & : & f < 2000Hz \\ 0.8f + 800 & : & f \geq 2000Hz \end{array} \right. \tag{8}$$

For the comparison the scaler function $T(f)$ is calculated for both warpings, and an interpolation of the mel error rates back to a linear scale has been performed.

The plot for the mel warping in Figure 4 shows a good similarity for the experimental and mathematical derived curves, except at

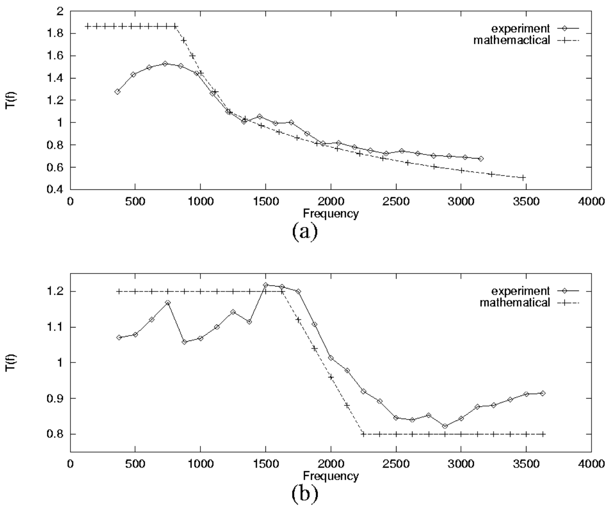

Figure 4: Comparision of experimental result and mathematical description for weighting function $T(f)$ a) mel-scale warping f_{mel}, equation (5), b) 2 component warping function f_{2lin}, equation (8)

low frequencies. For the f_{2lin} case the mathematical description and the experimental results show a very good correspondence with essentially no offset.

3.1. An estimation of a mel-like warping function

It has been shown that a transformation exists allowing the calculation of the error profile of another filter characteristic with the knowledge of the linear error profile and the transformation function. The goal now is to obtain a warping function which results in an equalised error profile from a suitable $T(f)$

$$T(f) = \frac{k}{e_{lin}(f)} \tag{9}$$

As a first approximation a mel-like warping is used:

$$f_T = \begin{cases} kf & : \ f < F \\ a\log(f-b)+d & : \ f \geq F \end{cases} \tag{10}$$

The frequency F of the function and its differentiation is considered to be continuous. The maximum output of the warping function at a frequency of 4000 Hz should also be 4000.

$$k = a\frac{\log e}{F-b} \tag{11}$$

$$kF = a\log(F-b)+d \tag{12}$$

Observing the graph for the linear and the mel error profile, there is an intersection of these two curves at approximately 1500 Hz and the number of filters below and above 1500 Hz is coincidently the same. Therefore the warped frequency range is split into two equal parts (0 to 2000 Hz and 2000 Hz to 4000 Hz). Thus

$$kF = a\frac{\log(4000-b)}{(F-b)} \qquad with\ F = 1500Hz. \tag{13}$$

With these equations the parameters in equation (10) can be evaluated to give the new mel-like warping function f_T in equation (14).

$$f_T = \begin{cases} \frac{4}{3}f & : \ f < 1500Hz \\ 4912\log(f+100) - 13738 & : \ f \geq 1500Hz \end{cases} \tag{14}$$

4. Recognition Results for f_{mel} and f_T

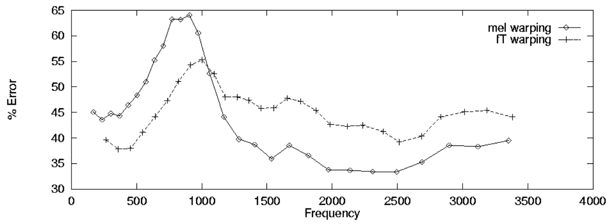

Figure 5: Recognition error profile for speaker set 1

Figure 5 shows error profiles directly equivalent to those in Figure 2, but for f_T and f_{mel}. Clearly f_T leads to a flatter profile, although perhaps not suprisingly the peak around 700-900 Hz remains.

4.1. Splitting into sub-bands

Next, we split the 32-bin frequency band into 2 and then 4 non-overlapping, equal, sub-bands and again perform sub-band recognition with each band in isolation.

Two sets of results are given in the tables. The first set, "set1", relates to 20 male speakers used in the initial experiments, and to derive the new warping function f_T above. The second set is a cross-validation set of 20 different speakers, again all-male.

Warp	Band	Band1	Band2	Band3	Band4	Combined
f_{mel}	1	3.27				3.27
	2	14.37		7.37		2.50
	4	30.53	41.40	21.67	22.73	3.63
f_T	1	3.83				3.83
	2	10.17		9.60		2.90
	4	23.57	33.57	25.93	27.60	2.87

(Speaker Set 1)

Warp	Band	Band1	Band2	Band3	Band4	Combined
f_{mel}	1	3.73				3.73
	2	15.87		7.03		3.07
	4	27.77	44.00	22.30	21.97	3.80
f_T	1	3.40				3.40
	2	11.40		8.90		3.20
	4	24.90	34.60	27.17	24.97	3.63

(Speaker Set 2)

Table 1: Recognition error for sub-bands, before and after recombination

Table 1 shows the recognition results for the sub-bands and after recombination with a linear weighted summation for the sub-band distances. For our tests equal weightings of these distances were used to produce an overall distance for the final decision.

For the two sub-band case, the sub-bands have similar recognition errors (10.17% and 9.6%). On the four sub-band case, the errors are not quite so similar, but nontheless are are more evenly distribution than for the f_{mel} case.

After recombination, error rates are slightly improved over the conventional one-band approach - a characteristic we find across numerous conditions.

4.2. Frequency warping and female speaker

In the previous tests we examined different frequency warping on a male speaker set. As known the mel warping was estimated on a male population. Our new warping function f_T was also developed and optimized on a male speaker set. In the next tests we will obtain the effect of these two warping functions on a female speaker set with 13 speaker.

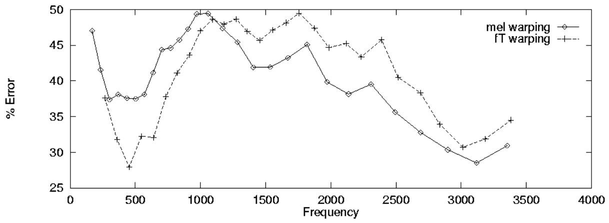

Figure 6: Recognition error profile for female speaker set

In Figure 6 it is conspicuous, that the peak error is shifted to higher frequencies. It can also be seen that the range with high error for this peak is broader than on a male speaker set.

The shifting of the peak error is explainable with the higher pitch of the female speaker. The broader error peak can be achieved with the harmonics of the pitch. When a higher pitch frequency was used, the harmonic frequencies are multiple of the pitch. Therefore the broader peak occur. It is obvious, that for a female speaker set, another warping should be used to get equalized errors.

Warp	Band	Band1	Band2	Band3	Band4	Combined
f_{mel}	1	7.33				7.33
	2	20.72		6.62		5.23
	4	32.62	43.23	49.85	56.92	7.28
f_T	1	4.82				4.82
	2	15.54		7.69		4.31
	4	25.90	30.62	30.51	19.38	5.08

Table 2: Recognition errors for a female speaker set

If we look on the sub-band errors for different number of sub-bands (Table 2), the equalized warping function shows a smaller variation for the error scores. Therefore the equalization works for female speaker too. It is unexpected, that the error after recombination is significantly lower for the new warping f_T than on a mel warping.

5. COMMENTS AND CONCLUSION

The characteristics of sub-band processing in the context of speaker recognition have been demonstrated and, in so doing, the mel and linear frequency scales have been shown to be sup-optimal. For example, both exhibit high error-rates just below 1000 Hz.

A new approach, based on error profiles, is proposed for deriving alternative warping functions, and a preliminary example (f_T) is shown to go some way to equalizing error rates across sub-bands, without degrading overall performance. More sophisticated forms of f_T are likely to further improve performance, for example in the region of 1000 Hz where error rates remain high - see Band 2 in Table 1.

Results for the cross-validation set (Set 2) mirror very well the results for the original speaker set. It is interesting to note that in all such experiments it is found that 2 sub-bands when recombined out-perform the conventional, single band case.

Test with an all-female speaker set gives a much lower overall error with the new warping function compared with the standard mel (4.8% cf 7.3%), although the error distribution across the sub-bands requires further flattening for the female set.

6. REFERENCES

1. H. Hermansky, S. Tibrewala, and M. Pavel. Towards ASR on partially corrupted speech. In *Proc. ICSLP-96*, 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-96*, volume 1, pages 426–429, 1996.
3. L. Besacier and J. Bonastre. Subband approach for automatic speaker recognition. In *Proc. AVBPA-97*, pages 195–202, 1997.
4. J. B. Allen. How do humans recognize speech? *IEEE Trans. Speech and Audio Processing*, 2(4):pages 567–577, October 1994.
5. L. Xu and J.S. Mason. Optimization of perceptually-based spectral transforms in speaker identification. In *Proc. Eurospeech-91*, pages 439–442, 1991.