Wavelet-Like Regression Features in the Cepstral Domain for Speaker Recognition

Jonathan Hume

Department of Electrical & Electronic Engineering,
University of Wales Swansea,
SWANSEA, SA2 8PP, UK.
email: J.Hume@swansea.ac.uk

ABSTRACT

This paper investigates the effects of using multiple time intervals for the calculation of regression coefficients. The technique that we have used is referred to as Wavelet-Like regression (WLR). Using this approach we have found that the underlying time series in the cepstral domain differs slightly depending upon the index of the series, and that by employing a technique that accounts for this, such as WLR, we may achieve an incremental improvement in recognition performance, at negligble extra costs.

1. INTRODUCTION

Temporal derivatives, when used in addition to existing static features have been found to give performance improvements for speaker recognition [1], [2], [3] and [4]. Similar results have also been reported for speech recognition [5], [6] and [7]. In speech recognition the dynamics are thought to be theoretically important in terms of compensating for spectral undershoot [8] and temporal spectral masking [9], whilst in speaker recognition the role of dynamic information is less well understood.

The derivatives can be directly calculated by differencing, or they may be approximated by a regression fit. The regression approach is reported as having a slight performance advantage [2], [5], and this is the approach that has been adopted here.

The interval over which the derivatives are taken is usually identical for each component of the derivative, although experiments that calculate the same derivative twice, using two different window lengths have been tried for speech recognition [6].

Speaker recognition work has suggested windows of between 90ms and 150ms for the first derivative [1],[2],[4], about 250ms for the second derivative [4] and 350ms for the third derivative [4]. It has also been reported that the use of the zeroth regression coefficient yields no performance improvement over the existing static features [4]. Turning to speech recognition, similar results have also been reported for the same derivatives [5], [10] [6].

In both speaker and speech recognition, some of the same pieces of work have also shown that very long window lengths, e.g. over 150ms for the first regression coefficient, give a slightly better performance than the normally used intervals, but these durations are typically rejected on practical grounds in favour of shorter effective windows [1], [2], [5], [10]. One notable case in particular, is Hanson and Applebaum, who have published extensively in this area [7], [12], [5], [10]. In [5], they find that optimum performance is achieved using 210ms for the first derivative, and over 300ms and 400ms for the second and third derivatives respectively. They do not use these optima, because they are longer than average syllable durations, and they assert that averaging over such a long time interval will not work so well on a more confusable vocabulary.

In [4], it is shown that the optimum dynamic interval is largely, but not completly, independent of the initial feature order. The observation that the optimum does exhibit a small dependency upon the number of components in the static vector, with generally low order vectors having slightly longer optima than the higher orders, forms the motivation for this work.

1.1. End-Effects

An implementation problem associated with the use of a moving window process on small amounts of data, is what to do at the ends. With short utterances, such as isolated words, the stage can be quickly reached where the number of dynamic vectors is significantly less than that of the original vectors; which can potentially result in useful information not being adequately represented. Padding methods that can be used to compensate for this include: zero [7], noise [7] and cyclic [11]; we have examined some of these in our initial experiments.

1.2. Wavelet-Like Regression

Regression fitting as a means of representing the dynamic information in the speech signal was first proposed for by Furui [1] for speaker recognition. The first-order regression coefficient is most commonly used, an equation for which is shown in Eqn. 1; the others may be found in [12].

$$R_{1k}(t, N, \Delta T) = \frac{\sum_{X=-(N-1)/2}^{(N-1)/2} X C_k(t + X\Delta T)}{\sum_{X=-(N-1)/2}^{(N-1)/2} X^2} \quad (1)$$

where N is the window length, C_k is the kth cepstral component and ΔT is the sampling period.

It has been suggested that the reason for the small performance advantage of regression, when compared to simple differencing, is due to its ability to cope better with the inherently noisy nature of the derivatives [2]. In the case of the first order derivative shown in Eqn. 1, it can be seen that it does this by constructing a weighted average about the window'ss central location ($X = 0$). From this it is obvious that the selection of the optimum window length(s) N for the regression calculation R_{rk}, is a balance between averaging out distortion, whilst still retaining resolution at a scale commensurate with the acoustic events under study.

In previous implementations of dynamic features, window lengths used have been constant for each component k of the regression vector. This assumes that each of the time series formed from the static vectors, have identical dynamic characteristics, and are equally effected by any noise present in the input signal. Our Wavelet-Like Regression uses the same regression equations, but abandons the above assumptions by using different window lengths for each componenent in the following way:

1. set the first and last regression window lengths to the desired values - when they are equal we have standard regression analysis.
2. linearly interpolate the intervals between these two extremes for each intervening coefficient.
3. quantise the values derived from 2) to the nearest (odd) window length.

The resultant feature we call a Wavelet-Like Regression (WLR) feature, in order to distinguish it from conventional regression features.

2. EXPERIMENTAL DETAILS

The database used in these experiments is the BT Millar, isolated-word speaker verification, digit database, down sampled to 8kHz and quantised to 16 bits. Each one of the digits is repeated 25 times across 5 sessions by each of the speakers.

Unless otherwise stated the following conditions prevail:

- 20 speaker all-male subset, using the first 3 versions for training and the last 15 for testing.
- static features, 14^{th} order mel cepstra (MC) calculated using 24 mel filters with a Hamming window of 32ms and a window overlap of 50%, discarding C_0.
- speaker models, text-dependent codebooks of size 16, derived using the binary-split, LBG algorithm.
- closed-set classification according to accumulated minimum distance criteria.

Benchmark results for inverse variance weighted MC (MCI) using the above conditions give an error rate of 11.37%.

2.1. End-Effects Experiments

Two possible ways for dealing with the end effects associated with dynamic features have been investigated, namely zero and cyclic (noise padding was not used, since it has already been shown to be inferior to zero padding [7]). The different schemes were compared with no padding, using the first order derivative derived from MC.

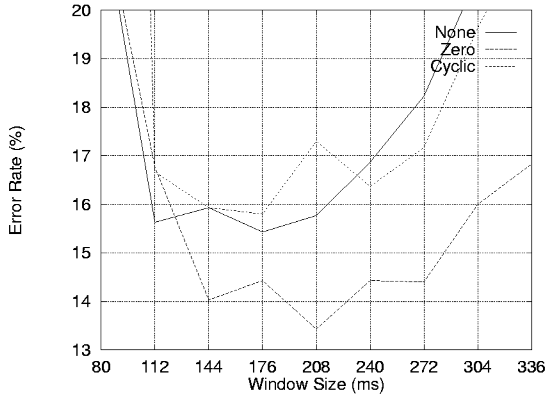

Figure 1: The effects of different padding schemes on recognition using R_1

The results shown in Fig 1, show that essentially our three options exhibit the same trends. With only zero padding showing a consistent performance advantage compared to the other two techniques. Both of the other approaches suggest windowing intervals that are shorter than that for zero padding. This indicates that zero padding deals with end-effects better than the other two approaches. Bearing these two points in mind we have adopted zero padding as our standard for the remaining regression experiments.

One unexpected observation from this experiment, is how well the zero padded first-order regression coefficients work. Error rates for windows between 144ms and 272ms are similar to those of the benchmark static features, even though the regression features were not inverse variance weighted. All of the previous work in speaker recognition has not shown this to be the case and most reported speech recognition results. By a process of elimination we suspect this difference is due to our use of zero padding and text dependent testing (previous work has either been text dependent unpadded [1], text independent unpadded [2], or text independent cyclic regression [4]. Interestingly, one of the papers that does show the first derivative with approximately equivalent performance to the static features, is a paper on speaker independent word recognition [5]. In that paper performance for the first derivatives is shown as being good between about 90ms and 250ms in normal conditions, with an optimum in good agreement with ours at 210ms.

2.2. Initial Wavelet-Like Regression Experiments

In these experiments we wanted to determine if using WLR is a sensible idea. In order to do this we have used WLR_1 features which have been derived from MC using zero padding.[1] The results of which, along with the performance of the optimal R_1 features are shown in Table 1.

Input	MC	
Processing	R_1	WLR_1
Opt. C_1	208	272
Win. (ms) C_{14}	208	112
Error (%)	13.53	10.97

Table 1: An initial comparison of the performance of WLR_1 and conventional regression

The results in Table 1 shows that using WLR_1 gives a small performance advantage. However, because the use of WLR effects the variances of the each time series to differing degrees. Because of this, it is likely that at least some of the improvement in the recognition performance, is due to empirically optimising the contribution of each component of the WLR to the distortion measure. In order to avoid this problem in the future, all the remaining experiments use inverse variance weighting.

2.3. Wavelet-Like Regression in the Cepstral Domain

In this set of experiments we determine the optimum calculation intervals for WLR_1, WLR_2 and WLR_3, these are compared with MCI and the optimum conventional regression coefficients. The standard results are shown in Table 2, whilst Table 3 shows the optimum results for WLR and Fig. 2 shows a typical example of an error contour map for WLR, in this case for WLR_1.

From Tables 2 and 3, we can see that the use of WLR gives a small performance advantage when compared to standard regression features.

Feature	MCI	R_1	R_2	R_3
Opt. Win. (ms)	NA	208	336	432
Error (%)	11.37	12.27	15.70	18.03

Table 2: Results for MCI and the conventional regression coefficients

Feature	WLR_1	WLR_2	WLR_3
Opt. Win. C_1	336	336	528
(ms) C_{14}	80	208	304
Error (%)	10.83	15.20	17.00

Table 3: Results for optimum WLR coefficients

Given the nature of the performance enhancement, it is important that the result is not data dependent. Figure 2 shows a contour map of the errors for WLR_1, verus the different window intervals for the time series C_1 and C_{14}, and helps us to see that generally lower errors occur to the right of the line depicting R_1 results. This region corresponds to generally better performance being achieved by using longer windows for the lower order regression components, and shorter ones for the higher parts. This generality implies that the use of WLR in order to enhance recogniton rates should not be very data dependent.

We have further tested the degree of data dependency of these results, by carrying out a cross validation test using the optimum regression intervals shown in Table 2 for R_1 and Table 4 for WLR_1. The conditions for this test set are identical to those previously used, except that the speakers used are 15 females, rather than the original 20 males. These give the results in Table 4.

Feature	MC	R_1	WLR_1
Error %	10.62	10.36	10.13

Table 4: WLR_1 cross validation, window lengths as in table 3

The cross-validation results above confirm our conclusion that the advantage obtained by using WLR_1, although small is real and that the same conclusion is likely to be true for the other WLR coefficients. We may also conclude that the dynamic characteristics of the initial static cepstral representation are not identical, and thus the interval over which the regression is calculated should be different depending upon the index of that component. The decreasing regression window length, for each successive cepstral time series, is also in agreement with what would be expected from previous work in [4].

[1] The use of MCI will give different results to those for MC

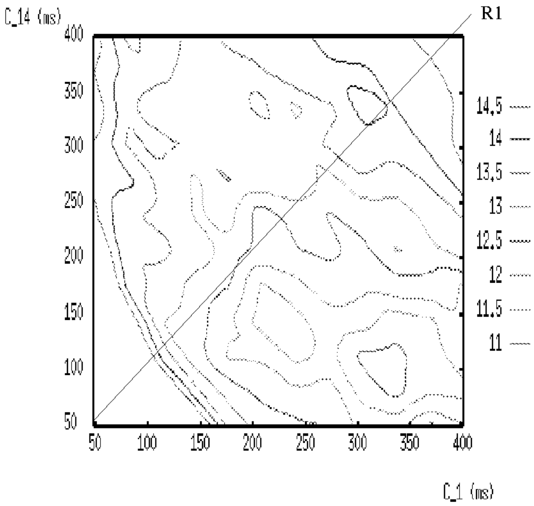

Figure 2: Contour map of the error rates for WLR_1, using a range of values for C_1 and C_14

3. COMMENTS AND CONCLUSIONS

The following two points may be drawn from this piece of work:

1. care must be taken in the optimisation of any moving window feature extraction technique, in order that the end-effects, accentuated by limited amounts of training data, are adequately dealt with. We have found that the best approach for dealing with this is zero padding, in agreement with [5].
2. the dynamic characteristics are not identical for each component of the cepstral vector. When a technique that accounts for this is used, such as WLR, then a small increase in performance is generally likely to be obtained. In particular we have found that the windows applied to the higher order cepstral coefficients should be shorter than those applied to to the lower, which is in agreement with what would be expected from previous published work in [4].

4. REFERENCES

1. S. Furui. Cepstral analysis techniques for automatic speaker verification. *IEEE Trans. ASSP-29*, 29:254–272, 1981.
2. F. K. Soong and A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. In *Proc. ICASSP-86, Tokgo*, volume 2, pages 877–880, April 1986.
3. L. Xu and J. S. Mason. Instantaneous and transitional perceptually-based features in speaker identification. In *Proc. Eurospeech-89*, pages 271–274, Paris, September 1989.
4. J. S. Mason and X. Zhang. Velocity and acceleration features in speaker recognition. In *Proc. ICASSP-91, Toronto, Canada*, volume 5, pages 3673–3677, 1991.
5. T. Applebaum and B. Hanson. Robust speaker-independent word recognition using spectral smoothing and temporal derivative features. In *Proc. EUSIPCO-90*, 1990.
6. S. Furui. On the use of hierarchal spectral dynamics in speech recognition. In *Proc. ICASSP-90*, pages 789–792, 1990.
7. B. A. Hanson and T. H. Applebaum. Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with lombard and noisy speech. In *Proc. ICASSP-90*, pages 857–860, 1990.
8. S. Furui. Speaker independent isolated word recognition based on emphasized spectral dynamics. In *Proc. ICASSP-86*, pages 37.10.1 – 37.10.4, 1986.
9. B. Strope and A. Alwan. A model of dynamic auditory perception and its application to robust speech recognition. In *Proc. ICASSP-96*, pages 37–40, 1996.
10. T. Applebaum and B. Hanson. Speaker independent recognition of noisy and Lombard speech. In *Proc. 120th meeting of the Acoustical Society of America*, San Diego, California, November 1990.
11. X. Zhang, J. S. Mason, and E. C. Andrews. Multiple dynamic features to enhance neural net based speaker verification. In *Proc. Eurospeech-91*, volume 2, pages 1411–1414, 1991.
12. T. H. Applebaum and B. A. Hanson. Tradeoffs in the design of regression features for word recognition. In *Proc. Eurospeech-91*, volume 3, pages 1203–1206, 1991.