EVALUATING FEATURE SET PERFORMANCE USING THE F-RATIO AND J-MEASURES

Simon Nicholson, Ben Milner** and Stephen Cox**

*University of East Anglia, Norwich, Norfolk, UK

**BT Laboratories, Martlesham Heath, Suffolk, UK.

ben@saltfarm.bt.co.uk sjc@sys.uea.ac.uk

ABSTRACT

Several methods of measuring the class separability in a feature space used to model speech sounds are described. A simple one-dimensional feature space is considered first where class discrimination is measured using the F-ratio. Using a conventional feature set comprising static, velocity and acceleration MFCCs a ranking of the discriminative ability of each coefficient is made for both a digit and alphabet vocabulary. These rankings are shown to be quite similar for the two vocabularies.

Discrimination measures are extended to multi-dimensional feature spaces using the J-measures. It is postulated that high correlation exists between feature sets which have a good measured class discrimination and those which give good recognition accuracy. Experiments are presented which measure this correlation and use it to predict recognition accuracy for a given set of features. These estimates are shown to be accurate for previously unseen combinations of features.

A brief analysis of the effect linear discriminant analysis on the feature space is made using these measures of separability. It is shown that LDA and separability measures are closely linked.

1. INTRODUCTION

A statistical pattern-matching system often functions by estimating parametric models of the distributions of the pattern classes in a feature space and then estimating the likelihood of an unknown pattern being produced by each distribution. In speech recognition these models are typically hidden Markov models (HMMs) which use multi-variate Gaussian probability density functions (PDFs) to represent the area in the feature space occupied by a particular speech class [1]. Each of these Gaussian PDFs is characterised by a mean vector and covariance matrix which are estimated from a set of training data. The feature space will be populated by a number of these distributions (determined by the number of speech classes). The accuracy of the recognition system is closely related to the arrangement of these distributions within the feature space. If the distributions are well separated, the recognition accuracy should be good, but overlap between distributions reduces the recognition accuracy.

Techniques such as linear discriminant analysis (LDA) [2] make use of this within-class and between-class covariance information to transform the feature space into a more discriminant sub-space. This work looks at ways of predicting feature space performance based on a knowledge of the class separability within that space. This has uses in selecting an optimal feature sub-set based on a trade-off between recognition performance and feature space dimensionality.

2. METHODS OF DISCRIMINATION

The ability of a feature to distinguish between two classes depends on both the distance between the two classes and the amount of scatter within the classes. A reasonable measure of class discrimination must take into account both the mean and variance of the classes. One such measure of separability between two classes is Fisher's discriminant [3].

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \qquad (1)$$

Where μ_1 and μ_2 are the two means or centroids of the classes and $\sigma_1{}^2$ and $\sigma_2{}^2$ are the variances of the classes. Higher class discrimination is measured when the class means are further apart or when the spread of the classes is smaller. Both these contribute to increasing the overall class separation.

Fisher's discriminant is able to measure the separability which exists between just two classes. For most tasks there are considerably more than just two classes. The F-ratio is an extension of Fisher's discriminant which provides a measure of separability between multiple classes.

$$F-ratio = \frac{\text{Variance of means (between-class)}}{\text{Mean of variances (within-class)}} \qquad (2)$$

Clearly if the spread of class means increases, or the distributions themselves become narrower then the separability will increase. A common method for modelling speech is to use a set of HMMs. As these are statistical models they form ideal candidates for extracting the between-class and within-class covariances for the given feature space. A class is defined to be the state of an HMM; hence state covariances provide within-class information and state means the between-class covariance.

To illustrate the F-ratio a set of digit-based full covariance HMMs is used to provide the within and between class covariances for a 27-D feature set comprising 9 static, 9 velocity and 9 acceleration MFCCs. Each state of each model is treated as a separate speech class. The F-ratio of each coefficient is shown in figure 1 as a solid line. To compare the F-ratios of the same 27-D feature set with a different vocabulary, a set of 26 alphabet HMMs are trained. The F-ratios of the 27 dimensions using the alphabet vocabulary is shown in figure 1 as a dotted line. To highlight the coefficients with lower F-ratios, log F-ratio is plotted.

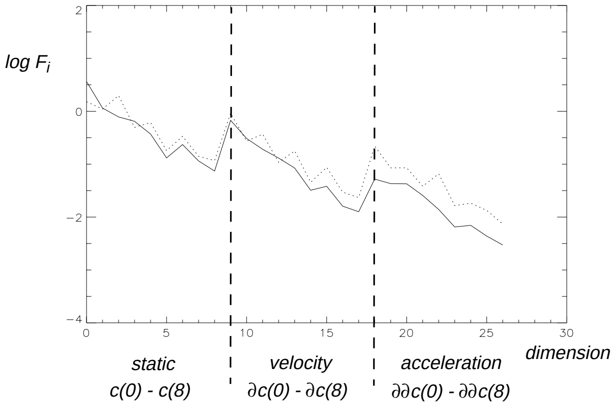


Figure 1 : F-ratio's of a 27-D MFCC feature vector based on a digit vocabulary (solid) and an alphabet vocabulary (dotted).

The graph shows three distinct regions which characterise the three different temporal regions of the feature vector - static, velocity and acceleration. Each region shows that lower quefrency coefficients generally have higher F-ratios and should therefore offer better class separation. It is also clear that static coefficients generally have higher discrimination than velocity coefficients, followed by acceleration coefficients. The F-ratio curve of the digit database is much smoother than that of the alphabet. This is attributed to each digit model having approximately three times more training data than the alphabet models - resulting in a more smooth estimate of the F-ratio.

Table 1 shows the ranked discriminative importance of each dimension in the feature vector for the two vocabularies by ordering each coefficient according to its F-ratio value. The rankings for the two vocabularies are similar - with the set of top 9 coefficients being the same for both vocabularies, although the ordering is slightly different.

Rank	F_{digit}	F_{alph}	Rank	F_{digit}	F_{alph}
1	c_0	c_2	15	$\partial\partial c_0$	∂c_3
2	c_1	c_0	16	$\partial\partial c_1$	∂c_6
3	c_2	c_1	17	$\partial\partial c_2$	$\partial\partial c_2$
4	∂c_0	∂c_0	18	∂c_6	$\partial\partial c_1$
5	c_3	c_4	19	∂c_5	$\partial\partial c_4$
6	c_4	c_3	20	$\partial\partial c_3$	∂c_5
7	∂c_1	∂c_2	21	∂c_7	$\partial\partial c_3$
8	c_6	c_6	22	$\partial\partial c_4$	∂c_7
9	∂c_2	∂c_1	23	∂c_8	∂c_8
10	∂c_3	$\partial\partial c_0$	24	$\partial\partial c_6$	$\partial\partial c_6$
11	c_5	c_5	25	$\partial\partial c_5$	$\partial\partial c_5$
12	c_7	∂c_4	26	$\partial\partial c_7$	$\partial\partial c_7$
13	∂c_4	c_7	27	$\partial\partial c_8$	$\partial\partial c_8$
14	c_8	c_8			

Table 1: Rank ordering of 27-D feature vector using F-ratios computed from digit vocabulary and alphabet vocabulary.

Other work has produced similar rankings of MFCC-based speech features - [5][6].

The F-ratio measures the separability of a single coefficient or dimension of the feature vector. To evaluate the discrimination of an entire feature set a multi-variate extension to the F-ratio is needed. A selection of such techniques are the J-measures, [4]. Four of these are shown below. The operator *tr(.)* is used to indicate the trace of a matrix and variable *D* is the feature space dimensionality.

$$J_1 = tr(\boldsymbol{W}^{-1}\boldsymbol{B}) \qquad (3)$$

$$J_2 = \ln\left(\frac{|\boldsymbol{B}|}{|\boldsymbol{W}|}\right) \qquad (4)$$

$$J_4 = \frac{tr(\boldsymbol{B})}{tr(\boldsymbol{W})} \qquad (5)$$

$$J_s = \sum_{i=1}^{D} \frac{b_{ii}}{w_{ii}} \qquad (6)$$

The J-measures take into account the locations of the classes using covariance information taken across all dimensions of the feature space. Matrix **B** is the between-class covariance, or covariance of class means, and measures how close the speech classes are from each another. Matrix **W** is the within-class covariance, or the average of the class covariances. This indicates how large the speech classes are. Both of these are can be computed directly from a set of full covariance trained HMMs.

To illustrate the usefulness of a measure of class separability a two-dimensional feature space with three speech classes is shown in figure 2. In dimension x_1 the class means are close together but the classes have narrow scatter. As shown this dimension gives good discrimination between the classes. Dimension x_2 has much wider spread of the class means, but has very wide class scatter. This makes discrimination along x_2 worse than along x_1, even though the class means are better separated. It is this within and between class covariance information that the J-measures exploit to measure the class separability of a feature space.

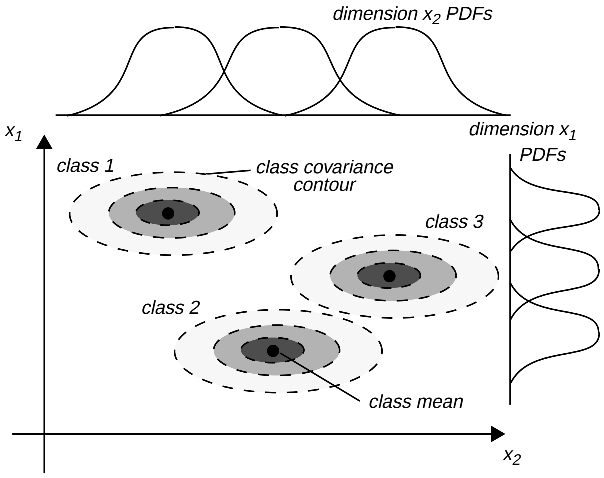
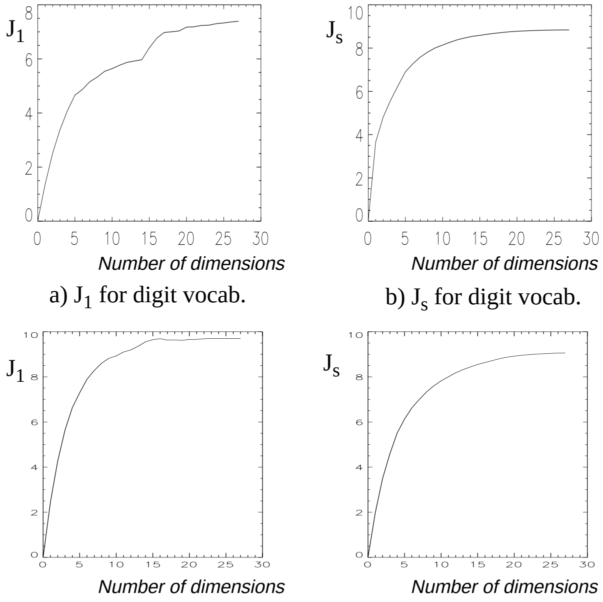


Figure 2 : Illustration of class separability in a two-dimensional feature space.

Using the same 27-D speech feature and the digit and alphabet vocabulary models described in section 1, J-measures can be computed.

Figure 3 shows the J_1 and J_s measures - figures 3a and 3b use the digit vocabulary and figures 3c and 3d use the alphabet vocabulary.

Following the rank ordering shown in table 1, those dimensions with lower F-ratios are removed first from the feature space and the J-measure of the new reduced dimensionality feature space re-calculated. The abscissa on the two graphs shows how many dimensions of the original 27-D feature space remain. The remaining dimensions are those which are measured to give better class discrimination.



a) J_1 for digit vocab.

b) J_s for digit vocab.

c) J_1 for alphabet vocab.

d) J_s for alphabet vocab.

Figure 3: Comparison of J_1 and J_s for digit and alphabet vocabularies using 27-D feature space.

J_1 and J_s both show that discrimination falls as increasing numbers of dimensions are removed from the feature space. Because the removal of dimensions is based on the F-ratio the first dimensions to be removed are those which give poorest discrimination. This results in the negligible drop in discrimination for a halving in feature space dimensionality.

3. RELATIONSHP OF LDA AND F-RATIO

The technique of LDA [2] is closely linked to the F-ratio and J-measure methods of determining separability between a set of classes. Assuming that the LDA transform matrix is given by **A**, then the within-class covariance of the LDA feature space is given by the identity matrix, i.e.

$$\mathbf{AWA}^T = \mathbf{I} \tag{7}$$

because of the rotation and scaling of the speech clusters [2]. Similarly the between-class covariance of the LDA feature space is a diagonal matrix,

$$\mathbf{ABA}^T = \Delta = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_D) \tag{8}$$

Considering equation (2) for computing the F-ratio shows that the diagonal elements of Δ (the eigenvalues) are in fact the F-ratios for the new feature space. From equations (3) and (6), and from the fact that the within-class covariance is now the identity matrix, both J_1 and J_s in the LDA feature space reduce to a simple summation of the F-ratios.

4. PREDICTING PERFORMANCE

The performance of a recognition system is closely linked with the choice of feature space. This implies that there is a close correlation between recognition accuracy and the separability of the speech classes as measured by the J-measures. Recognition accuracy based on the 27-D feature space with the digit-based vocabulary is shown in figure 4. Using the F-ratio rank ordering in table 1, dimensions with poor discriminative ability are successively removed from the speech feature. The recognition performance of this new feature sub-space is then re-evaluated experimentally. The number of dimensions remaining is indicated along the abscissa.

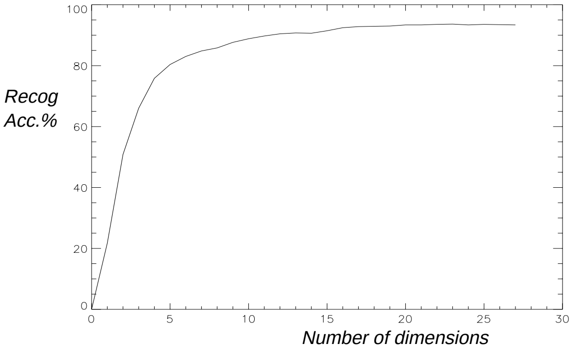
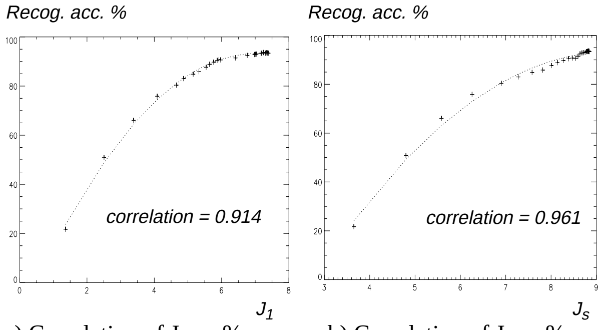


Figure 4 : Recognition accuracy of feature subsets.

Examination of figures 3a, 3b and 4 shows that J_1 and particularly J_s correlate closely to the recognition accuracy as the dimensionality of the feature space is successively reduced. This correlation is illustrated in the scatter plots of figure 5.



a) Correlation of J_1 vs %acc

b) Correlation of J_s vs %acc

Figure 5 : Correlation of J_1 and J_s with recognition accuracy

The correlation coefficients show that the J-measures correlate highly with recognition accuracy. In particular J_s slightly more than J_1.

Given the strong correlation between class discrimination and recognition performance, an estimate of recognition performance can be made from the J-measure. Using the scatter plots of figure 5, a second order polynomial can be estimated to best fit the points - these are shown by the dotted lines and are of the form

$$rec(J) = a_0 + a_1 J + a_2 J^2 \tag{9}$$

with the coefficients a_0, a_1 and a_2 estimated using linear regression. The two recognition estimates use J_1 and J_s respectively.

$$rec(J_1) = -13.83 + 30.46J_1 - 2.17J_1^2 \qquad (10)$$

$$rec(J_s) = -93.34 + 40.00J_s - 2.14J_s^2 \qquad (11)$$

5. PREDICTION RESULTS

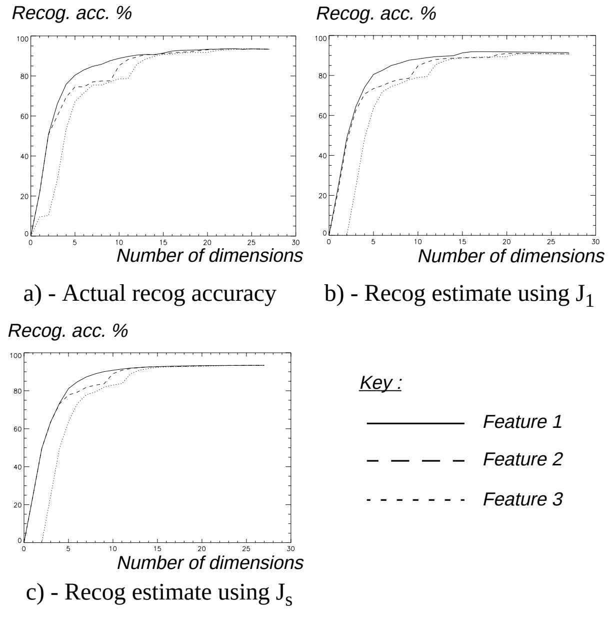
Using the recognition prediction equations it is possible to predict the recognition accuracy of a previously unseen feature space, based on its J-measure. In this experiment the feature is based on the same 27-D MFCCs, but the truncation is done in a different order than the F-ratio ranking. Three different features are produced which are derived from the original 27-D feature space.

Feature 1 - this feature space begins with the full 27-D space and truncates dimensions according to their F-ratios. This is the same feature as described by figure 4 and serves as a baseline test.

Feature 2 - this begins with the 27-D feature which is successively truncated by dropping coefficients from the end of the feature vector. i.e. dimensions are removed in the order $\partial\partial c_8$, $\partial\partial c_7$, $\partial\partial c_6$, etc., leaving just c_0.

Feature 3 - this the same as Feature 2, except that $\partial\partial c_8$ and $\partial\partial c_7$ are retained in the feature until the end of truncation, with dimension removal beginning at $\partial\partial c_6$, $\partial\partial c_5$, $\partial\partial c_4$, etc. With just three dimensions remaining c_0 is removed leaving just $\partial\partial c_8$ and $\partial\partial c_7$ as the feature vector. Finally $\partial\partial c_8$ remains as the single dimension.

Figure 6a shows the recognition performance of the three feature sets as a function of the feature space dimensionality. Figures 6b and 6c show the estimated recognition performance based on the J_1 and J_s measures using equations (10) and (11) respectively.



a) - Actual recog accuracy

b) - Recog estimate using J_1

c) - Recog estimate using J_s

Figure 6 : Actual and estimated recognition accuracy.

All the predicted recognition accuracies shown in figures 6b and 6c appear close to the actual recognition performances measured in figure 6a. Both the J_1 and J_s based recognition estimates predict well the decay of feature set performance as dimensionality is reduced. Feature truncation according to the F-ratio ranking sustains much higher performance than the arbitrary truncations of feature 2 and 3, with no drop in accuracy for a halving in feature space.

Considering feature 2, both the J_1 and J_s estimates correctly predict the sudden fall off in performance when the feature space is reduced to 9 dimensions. For feature 2 the drop to just 9 dimensions is the breakpoint where the feature comprises only the static cepstral coefficients.

Feature 3 exhibits a similar sudden drop in performance when the dimensionality is reduced to 11. At this stage the feature comprises only the static MFCCs plus the two acceleration coefficients $\partial\partial c_7$ and $\partial\partial c_8$. Both the J_1 and J_s performance estimates correctly predict this fall off. When only the two acceleration coefficients remain in feature 3 both the J_1 and J_s measures become very small. As shown this leads to an underestimation in performance. This can be attributed to the fact that no very small J-measures were used in the scatter plots of figure 5 leading to the inability of the prediction equations to cope with this.

6. CONCLUSION

Several methods of measuring the separability within a feature space have been described. These range from the one dimensional F-ratio to a selection of multi-dimensional J-measures. These measures of discrimination are shown to correlate highly with the recognition performance attained on an isolated digit task confirming the hypothesis that good class separability gives good recognition accuracy. Based on this correlation, predictions of the recognition performance of previously unseen feature spaces are made from the J-measures and are shown to be reasonably accurate. In particular results have shown that sudden drops in performance can be accurately predicted. Results have also shown that feature selection according to F-ratio rankings yields a good trade-off between recognition accuracy and dimensionality. In particular a halving in feature space dimensionality gave no reduction in accuracy.

7. REFERENCES

1. L.R. Rabiner, “A tutorial on hidden Markov models and selected applications for speech recognition”, Proc. IEEE, Vol. 77, No.2, pp 257-285, Feb. 1989.

2. E.S. Parris and M.J. Carey, “Estimating linear discriminant parameters for continuous density hidden Markov models”, Proc. ICSLP, pp. 215-218, 1994.

3. T. Parsons, Voice and Speech Processing, McGraw-Hill, 1987.

4. K. Fukunaga, Introduction to statistical pattern recognition, Academic Press, 1974.

5. K.K. Paliwal, “Dimensionality reduction of the enhanced feature set for the HMM-based speech recogniser”, Digital Signal Processing, vol. 2, pp. 157-173, 1992.

6. E.L. Bocchieri and J.G. Wilpon, “Discriminative feature selection for speech recognition”, Computer, Speech and Language, Vol. 7, pp. 229-246, 1993.