MACHINE LEARNING APPLIED TO THE SONIC CLASSIFICATION OF MUSICAL INSTRUMENT LOUDSPEAKERS

A J Harper Celestion / GP Acoustics (UK) Ltd.

1 INTRODUCTION

Machine learning is the field of study that gives computers the ability to learn, without being explicitly programmed. The more experience a program has, the better it gets at its task. In the case considered in this project, the more speakers that are measured, the better a program gets at accurately predicting a listener's subjective judgement.

Standardised measurement and processing techniques exist which indicate how a loudspeaker performs in one regard. Many of these relate well to subjective impression, but no single measurement can tell the whole story. Loudspeaker engineers learn to relate a range of measured information to the sound of loudspeakers over many years of experience, often knowing largely how a unit will sound before listening. This methodology replicates the learning element, allowing a program to find the optimal separation between groups of loudspeakers when fed with a range of the most meaningful measurements. Unclassified drive units can then be classified as good or bad in a meaningful way, with a quantifiable output. These classifications correlate highly with subjective judgements.

This work outlines the relevant concepts of machine learning in relation to loudspeaker classification, and gives an overview of three available methods, before outlining the reasons for the chosen solution. These techniques give an interesting insight into the relative importance of each measurement as an indicator of subjective judgement, and final results show a greatly improved separation of groupings compared to alternative techniques or any one measurement alone.

An efficient listening test methodology is described which is uniquely suitable for the purpose. This provided maximum audible difference between groups, while being repeatable, controlled and time efficient. Drive units which could be judged repeatedly with confidence were selected, and their measurements used to train, tune, and test the models.

It should be stressed that musical instrument speakers are intended to produce sound, rather than reproduce sound[1] and that inaccuracy of reproduction is the design intention. An electric guitar played through a hi-fi speaker, or recorded music played through a guitar speaker is an illuminating demonstration of this. In this context *good* refers to the ideal sonic character of that speaker for the typical applications where it is used. Results are not directly transferable to speakers intended to reproduce sound.

2 BACKGROUND

For musical instrument loudspeakers the desired tonal response is a fine balance of different characteristics. There is no pre-defined ideal for any one measurement, but rather a desired recipe of tonal ingredients. A little more of one attribute could be acceptable if there is a little less of another, for this reason a multi-dimensional approach was required.

Psychoacoustic metrics have been successfully applied as tonal descriptors and used to evaluate the sound quality of grand pianos[2,3]. Venezuela[3] combined sharpness with a weighted distribution of specific loudness in a tonal quality metric. This correctly ranked four pianos from the likes of Steinway and Bösendorfer, with the predicted quality ratings correlating highly to those made in subjective evaluations.

Fastl and Zwicker[4] put forward a model for sensory pleasantness that included the exponential combination of sharpness, roughness, tonality and loudness. Relative results correlate well with those based on listening tests for a range of sounds, which further emphasises the need to combine meaningful descriptors when predicting a subjective judgement.

It is common in psychoacoustics to apply statistical techniques not only to the analysis of subjective listening test data, but also to the objective measurements in an attempt to correlate the objective and subjective. Staffeldt[5] looked at high-quality loudspeakers and used dimensional analysis to map objective data from a wide range of measurements to two perceptual dimensions. A correlation was observed between both sound power response and the phase response, when compared to these perceptual dimensions. One dimension related to qualities concerning treble reproduction, and the other bass reproduction.

Klippel[6] used statistical analysis to extract seven features that described the listening impression of consumer audio speakers; this included a modified version of sharpness called treble stressing. Sophisticated psychoacoustic features were developed based on the anechoic and predicted in-room response, and the defect was calculated relative to the pre-determined ideal for each. The final quality metric was a weighted linear combination of three features; feeling of space, discolouration, and brightness.

Olive[7,8] built on the work of Toole[9,10] to correlate new spectral features to subjective preference ratings and applied these to a range of predicted responses calculated from anechoic measurements. Principal component analysis (PCA) was used to assess the features with the highest contribution to the variance seen, then again to project the 23 features onto a two-dimensional factor space. For the main model four features were selected on the basis that they correlated to the two perceptual dimensions, but not to each other. These features were then weighted in a linear model which was shown to correlate highly with subjective preference rating.

The above studies apply Mallow's C_p criterion to limit the number of features used, in order to prevent overfitting and allow the model to be applicable to other datasets. This paper discusses alternative techniques to prevent overfitting, the results of which don't suffer from the performance reduction that occurs with dimensionality reduction or the removal of independent, orthogonal features. This aim here was not to model or explain the perceptual process, but rather to build a working logistic model, with the highest possible correlation when applied to previously unseen data.

One important distinction is that the problem addressed here is one of classification, where discrete predicted judgements are required rather than a continuous predicted preference rating. It therefore is a problem of logistic rather than linear regression; to obtain the boundary between subjectively assessed groups of objective data, rather than the line of best fit between subjective and objective preference ratings. An empirical method is described, to find the optimal non-linear boundary between two groups of loudspeaker measurements, one which is automatically recalculated as more measurements are presented.

3 LISTENING TEST METHOD

Before any algorithm development could take place, there was a need to gather many speakers of each grouping. These units would be used to train, tune, and test the machine learning models. As each of these stages required a distinct batch of drive units, the number of units being assessed was large. There was a need to develop a repeatable and efficient test method.

3.1 Developing the Listening Test Method

We are fortunate at Celestion to have a state of the art bespoke listening room, as pictured in *Figure 1*.

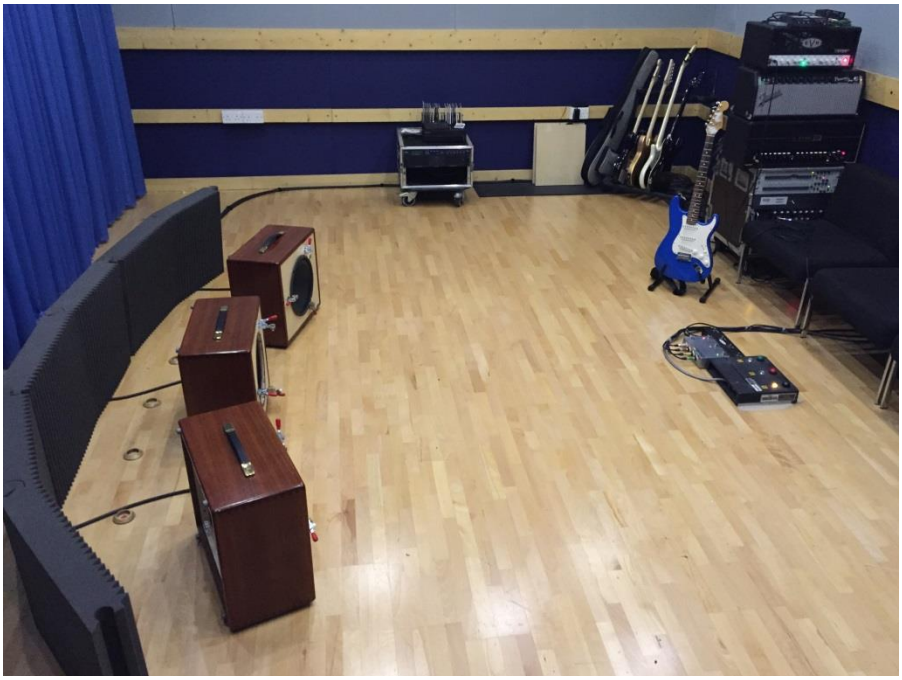

Figure 1. Celestion Listening Room

This was designed by Philip Newell to have the following key attributes which are vital for reliable listening test results:

- Minimal placement bias – such as differences in reflections / room interaction between left / right / centre.
- No significant room modes – so that differences at the listener locations are largely loudspeaker-related.
- Diffusive front wall, to help blend the off-axis sound at the listening position.
- Consistent reverberation time of around 150 ms from bass to high-midrange (120-4000 Hz).
- Floor coupling for a typical response.

This provided an excellent environment to judge drive units for preliminary investigations, and to amass a selection of suitable units for the main listening tests, described below.

Confidence was required that each unit used to train the model belonged to its assigned group, and therefore had reliable listening test results. To obtain a statistically significant judgement, with an alpha value of 0.05 for a two tailed test, a speaker must be given the same judgement in at least nine out of ten listening tests. Around a hundred grouped units were required; this meant that around twice this number were to be judged in total, each ten times for statistical significance. The

sheer number of required listening tests necessitated a listening method with no physical changeover of speakers between tests. It was therefore decided to take controlled recordings of each speaker, allowing simple switching between recorded samples of each test unit.
An automated test method was designed with a simple graphical user interface (GUI) that would randomise playback and automatically save results for statistical analysis, shown in *Figure 2*.

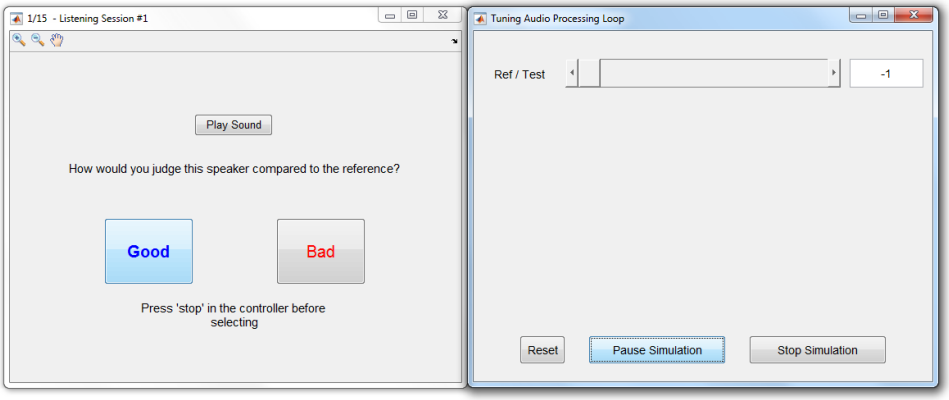

Figure 2. Listening Test GUI

After listening, units with statistically significant judgements were selected, and their measurements used to train, tune, and test the models.

Program material was selected to give maximum difference between units, and was controlled through re-amping of one source signal; a direct input (DI) electric guitar recording.

Notable psychoacoustic effects were observed throughout this process. Differences between units were found to be far more audible at very low playback SPLs, to the extent that barely audible differences with moderate SPLs under headphones became obvious in the headphone spill. This was attributed to widening of masking bands as SPL increases, and because of this, headphone level was drastically reduced and controlled to good effect[4].This technique allowed for high recording levels, where the amp and speaker are performing as intended, but controlled playback levels, where listening ability is maximised.

4 INPUTS – SONIC FEATURES

The machine learning program requires numeric inputs based on meaningful measurements and post-processing techniques. These quantifiable inputs will be referred to from here on in as features. It was of key importance that well-performing features were selected, and that these had a high probability of relating to the subjective judgements. Engineering knowledge was applied to ensure causation as well as correlation. Where applicable, mid-ear filtering and perceptual scaling were applied (conversion to sone scale of perceived loudness, and Bark scale of critical frequency bands) before relevant features were computed.

4.1 Feature Selection

While the number of features is not limited, it is good practice to only use those which provide a statistically significant separation individually. A full explanation of the features investigated and the reasons for the final choice is beyond the scope of this paper, however some of the highest performing and interesting features will be discussed.

T-tests were used to assess whether the groupings for each feature were sufficiently different from each other. The value for alpha of 0.05 was adopted as the standard due to the lack of preliminary data on which to base this decision. Features which return a p value of ($p<0.05$) were taken to be statistically significant, and were therefore included as features.
While the p value gives the probability that the separation seen in values for each feature occurred by chance, it gives no comparative information between features. In order to assess relative feature performance, the optimum point of separation between groups was calculated for each feature. This was done using a logistic regression classifier; using methods discussed in *Section 5*. Once this optimal boundary is obtained, the quantity of correctly and incorrectly classified units from each original group can be quantified. It is common to visualise the possible outcomes in a confusion matrix, as shown in *Table 1*.

		Predicted Classification	
		Good	**Bad**
Subjective Judgement	**Good**	True Positive	False Positive
	Bad	False Negative	True Negative

Table 1. Confusion Matrix

The relative proportions of each outcome in the confusion matrix allow various performance measures to be calculated, such as:

- Precision; the proportion good units that are correctly classified.
- Recall; of the units classified as good, this proportion was classified correctly.
- F1-Score; the harmonic mean of precision and recall.
- Mathew's Correlation Coefficient; a balanced measure of correlation for binary data.

These all relate in different ways to the relative proportions of outcomes in the confusion matrix shown. It should however be noted that the relative importance of each performance measure will depend on the application. These performance measures can be used to analyse the final algorithms, as well as individual features. Likewise, which performance measures are selected for the final tuning and optimisation will greatly affect the final performance of the program, and will produce classifiers with very different strengths.

In total 27 features performed well and were selected. These can loosely be divided into five categories:

- Psychoacoustic metrics[4], e.g. sharpness
- Electroacoustic parameters
- Energy features[11], e.g. energy centre
- Those relating to cone behaviour
- Spectral features, e.g. HF roll-off, brightness

These features act as the inputs to the machine learning algorithms, with each drive unit having a single value for each feature.

When any of the selected features are plotted individually two distinct groups can be seen, having approximately normal distributions about differing means. However, as can be seen in *Figure 3*, not only are the groups not linearly separable, but results overlap considerably. The probability density function of one of the highest performing features; brightness is plotted in *Figure 3*.

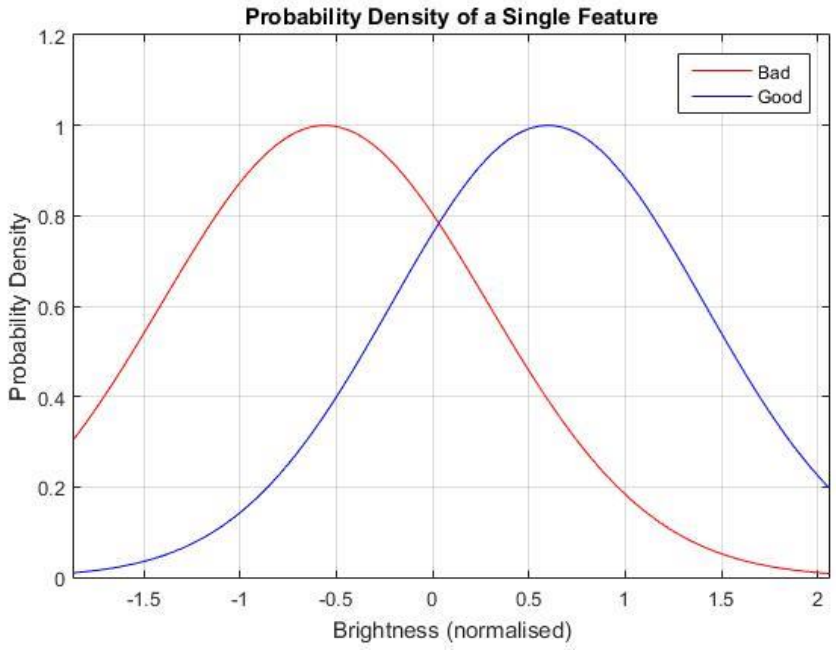

Figure 3. Probability Density Function for a Single High-Performing Feature

As you can see from *Figure 3*, whichever value of brightness is selected as the good/bad boundary, a significant proportion of units from each group will be incorrectly classified. The aim of combining many features into one overall decision with a machine learning classifier is to improve this separation. On balance, considering all features, it should be possible to make a much more confident decision on the classification for each loudspeaker.

Before the next stage each feature was normalised to its mean, and scaled to unit variance. This process helped ensure the optimisation algorithms progress towards the minimum at a rate that is equally biased for all features; preventing any overshooting and reducing processing time.

5 MACHINE LEARNING

5.1 Linear Logistic Regression

Each feature represents one dimension. Linear logistic regression was applied at first to each feature to assess relative performance, as described above and illustrated in *Figure 3.*

Linear logistic regression uses the process of minimising the cost function for the hypothesised boundary, in order to find the optimal separation between groups of measurements; the decision boundary. The parameters obtained by this method are fed into the hypothesis function, along with the quantified features, which then outputs the probability of a positive result[12].

For logistic regression, the hypothesis function is:

$$h_\theta = g(\theta^T X) \tag{1}$$

Where *X* is the feature matrix, θ^T is the transpose of the parameter vector; the coefficients for the linear boundary, and $g(z)$; the logistic function, is:

$$g(z) = \frac{1}{1+e^{-z}} \tag{2}$$

This logistic function was used as it is asymptotic to 0 for $(z) \ll 0$, and to 1 when $g(z) \gg 0$. Its output therefore tends to 0 for *bad*, or 1 for *good*, with a small proportion of values in between.

The hypothesis function represents the probability of the *good* outcome, given the features that are input as columns in X, and is parameterised by θ.

$$h_\theta = P(y = 1|X; \theta) \tag{3}$$

The cost function used for measurements of a single unit was:

$$Cost(h_\theta, x, y) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -\log(1 - h_\theta(x)) & if\ y = 0 \end{cases} \tag{4}$$

When the subjective judgement is y = 1 (training unit is *good*) and the hypothesis function predicts 0 (classification is *bad*), this cost tends to infinity, and to zero where it predicts 1. Likewise, for y = 0 (training unit is *bad*) this tends to infinity where the hypothesis function predicts 1, and zero when it predicts 0. This makes it ideal, as a large cost is paid for incorrect predictions.

This can be expressed for m units as:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} -y^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) - (1 - y^{(i)}) \log\left(1 - h_\theta(x^{(i)})\right) \tag{5}$$

This sums over all training units, and combines both cases for y, selecting each of the above cost function as appropriate[12].

The cost function $J(\theta)$ can then be minimised for θ using methods such as gradient decent, or by using optimisation functions available in software such as MATLAB®. The optimised parameter vector θ then gives the coefficients of the most suitable decision boundary.

The linear boundary obtained is relative to the number of features input; or the number of dimensions within which it is calculated. For one dimension, with one feature, this becomes a value. For two dimensions, as illustrated in *Figure 4*, this boundary becomes a line.

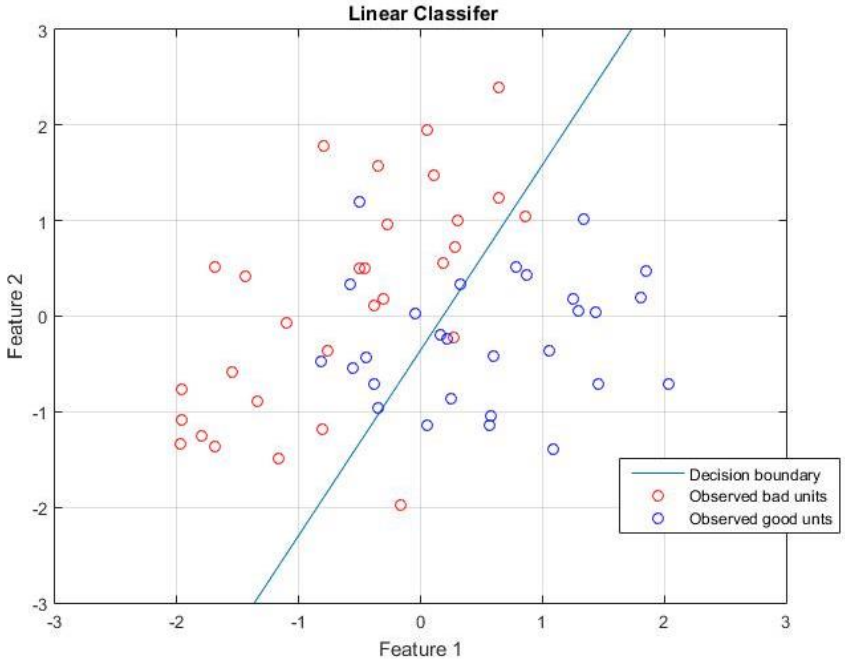

Figure 4. Relationship between two Features and the Linear Decision Boundary

For three dimensions, where three features are being evaluated together, this boundary becomes a plane; a two-dimensional plane in three-dimensional space, as illustrated in *Figure 5*. In general, the decision boundary is a sub-space, with dimension order one less than the space being evaluated.

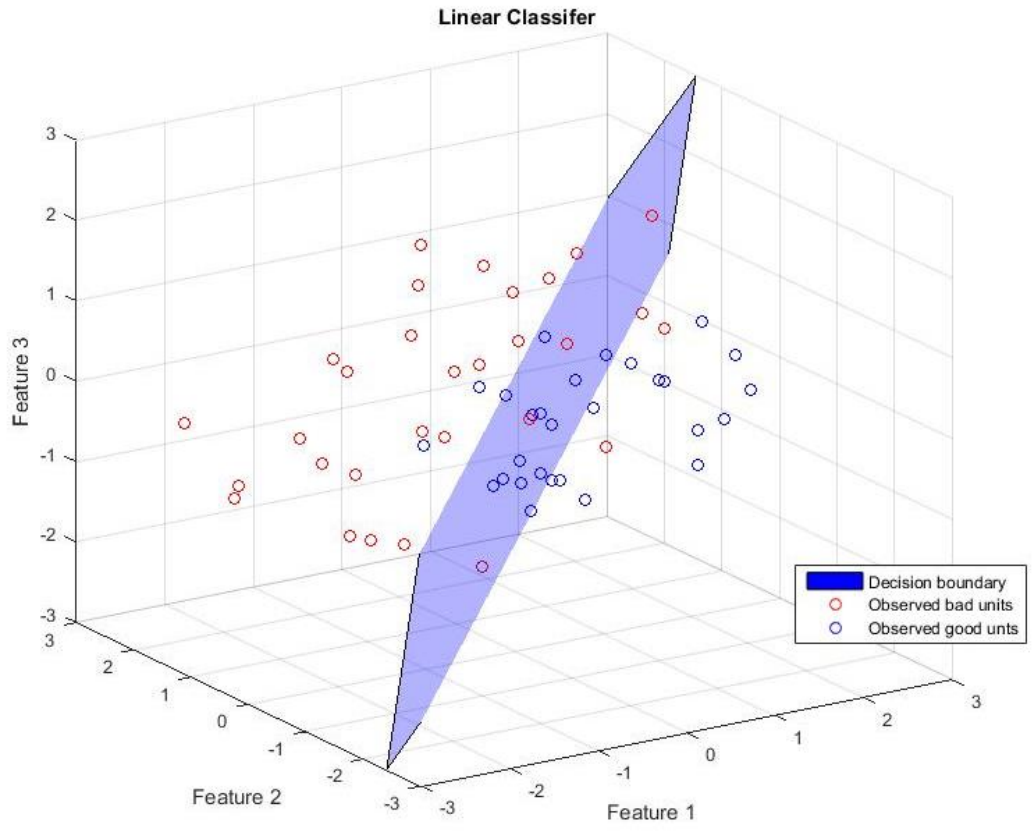

Figure 5. Comparison between Three Features and the Linear Decision Boundary

The optimised parameter vector θ for linear logistic regression can be used to analyse the relative contribution of each feature to the good/bad decision making process. This relative weighting for each feature, as shown in *Figure 6,* gives a detailed insight into the relative importance of each measurement and post-processing technique.

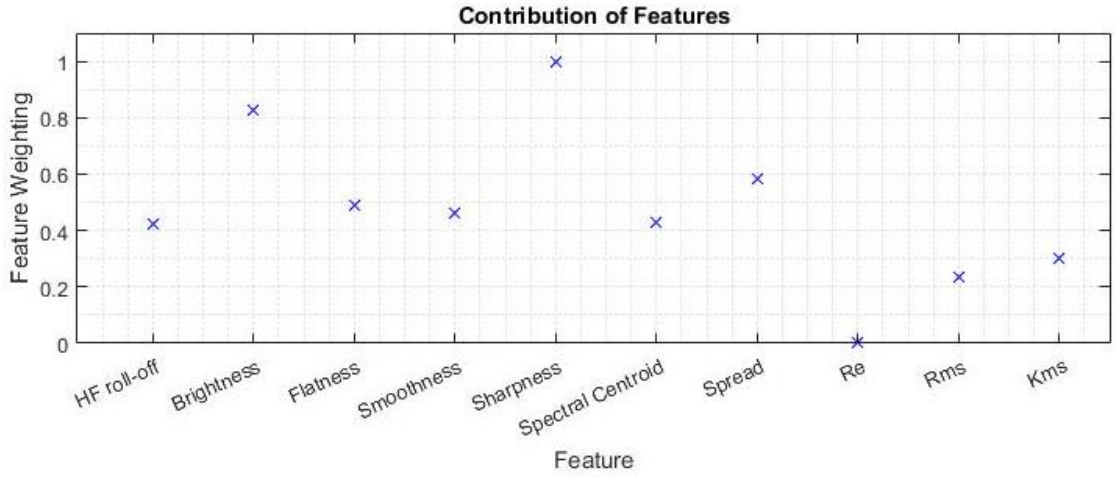

Figure 6. Calculated Weighting for a Selection of Features

As can be seen from *Figure 6,* in this instance sharpness is weighted as the most important feature to the separation, with brightness a close second. Re was found to be unimportant to the classification process, and was weighted to have near zero contribution. Rms and Kms were shown to be useful, but less so than the remaining spectral features.

When non-linear relationships exist between features, or higher-order boundaries are preferred in general, it is simple to create new features as products of existing features or higher order products of a single feature. These just become additional columns of the feature matrix and the above methods are still applicable. This was applied with reasonable success, improving performance relative to a linear boundary. However, more advanced techniques exist for this which further improved performance.

5.2 Support Vector Machines

Support vector machine (SVM) classifiers are an alternative to the above methods, and are often more efficient and easier to optimise[13]. SVM methods optimise the boundary to be as far away as possible from both groups of data points, and in their most basic linear form are simply more efficient methods of linear logistic regression. However, when used with suitable kernels, these can be used to find higher-dimensional boundaries. Kernels are algorithms which map the feature space to higher dimensions, allowing complex non-linear boundaries to be calculated. They do this in an efficient way by calculating the necessary inner products for combinations of higher order features, without ever calculating the exact coordinates in that space[13].

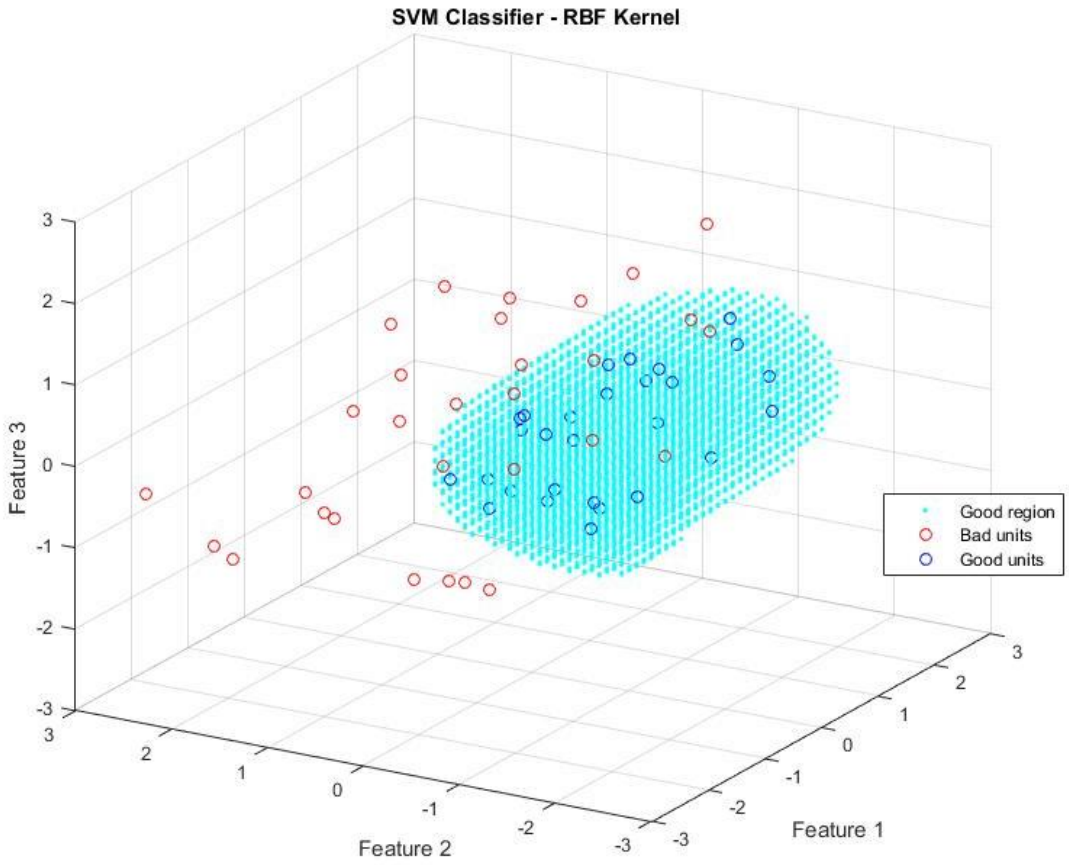

Figure 7. Comparison between Three Features and SVM Decision Boundary

Figure 7 illustrates the same data as *Figure 5*, but using SVM the boundary now becomes a hyperplane around the good units. Units outside this region are deemed bad but only the good region is shown for clarity.

5.3 Improving the Model – Tuning and Regularisation

Regularisation was applied to ensure the boundary was not oversimplified, so as to not be representative of the data trend, or overly complex, where it fits one data set precisely but is less applicable to other similar data sets.

5.3.1 Regularisation

The key concept of regularisation is how well the optimal boundary will translate to other datasets. If an overly complex model is used, correlation with data can be artificially high; when the classifier is applied to similar but non-identical measurements the error will be significantly larger. In order to avoid this, a separate cross-validation dataset is used, to tune the model and ensure it has not been made overly complex in order to fit the original data. *Figure 8* shows a typical plot used to select the regularisation parameter lambda, which is included in an addition term in the cost function.

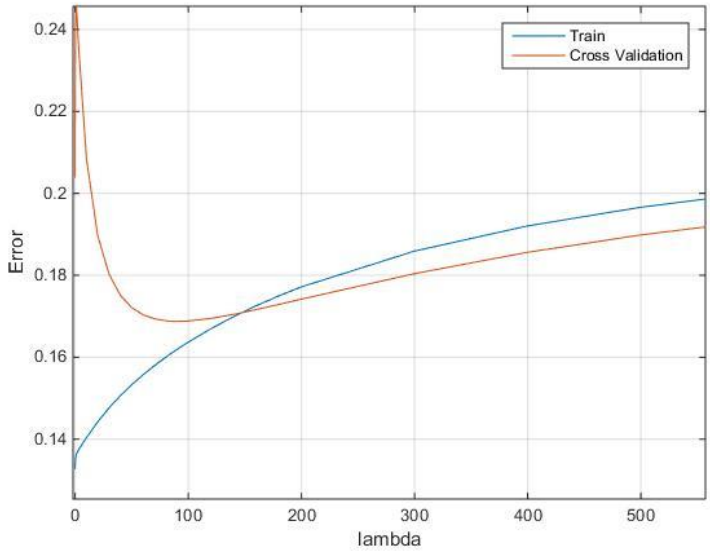

Figure 8. Error against Regularisation Parameter

The ideal classifier has the minimum error based on the cross-validation set. This increases the error for the training set, relative to no regularisation (lambda = 0), but the model will now be more transferable to other sets of measurements. With no regularisation the boundary can over-fit the specific training data, and won't generalise well. Regularisation was employed on every model, from linear logistic regression, through SVM, to ensemble methods outlined below.

To ensure reliable regularisation, the number of training units was kept far higher than the number of features. Otherwise points become perfectly separable by even linear classifiers, and some of the matrix transformations required for efficient optimisation will not work.

5.3.2 Dimensionality Reduction

Any features that are directly related to others will be redundant, and won't add any value. These will also cause the covariance matrix to be non-invertible, which rules out the most efficient optimisation methods. In this instance dimensionality reduction techniques such as principal component analysis (PCA) can be applied to project related features into a lower dimensional sub-space, reducing training time and storage space, and ensuring orthogonality of features. This technique has been applied in previous studies to allow many features to be projected onto a small number of key perceptual dimensions, but the logistic techniques used here are distinctly separate from this. The relatively small number of features used in this study compared to usual applications of machine learning did not make storage or speed a concern, and as the features chosen were orthogonal, dimensionality reduction would have only reduced performance.

5.4 Enhanced Techniques – Ensemble Methods and Cascade Architecture

5.4.1 Ensemble Methods

Ensemble methods were used to combine many reasonably-performing classifiers into one high-performing classifier. Some trained listeners use similar techniques; listening for several aspects of performance, each of which could be broken down into sub-criteria. For example the dynamics performance might be based on tightness and punchiness while the frequency response might be

based, amongst others, on frequency extension and spectral balance. Each of these general performance areas are then combined and weighted to make the overall judgement.

While ensemble methods further improve the performance of the model, one drawback is that it's difficult to dissect the information and relate it back to what makes a good speaker. In this respect the simpler algorithms taught us more about the ideal performance, despite ultimately performing poorly compared to SVM or ensemble methods.

5.4.2 Cascade Architecture

Cascade architecture can be used as a final improvement to the model. This technique was originally proposed to speed up real time facial recognition, by implementing a series of classifiers of increasing complexity, only using all available features when necessary[14].

As the number of features employed here was comparatively few at only 27, computing power was not a concern. The full featured and regularised ensemble classifier was used in each stage, with erroneously classified training examples removed between stages. While it could be argued that these outliers are important to the placement of the decision boundary, their removal gave a further marked improvement in performance on the test data.

This final method, taking input from all features, was capable of separating the units into two distinct groups, as illustrated by *Figure 9*, and outperformed both linear and non-linear logistic regression as well as SVM with a range of kernels.

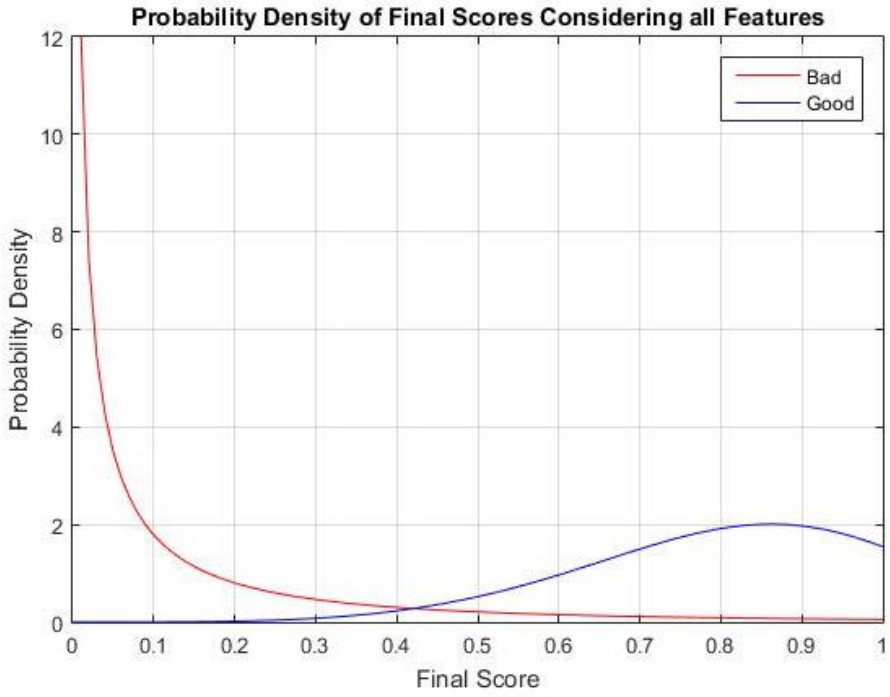

Figure 9. Probability Density Function of Final Algorithm

An improved separation can clearly be seen in results using the final algorithm, compared to that for any single high-performing feature, an example of which is shown in *Figure 3*. The proportion of units from each group that could be misclassified has been drastically reduced, as seen by the significantly reduced overlap between probability density functions.

6 CONCLUSION

Based on a handful of basic measurements, it has been shown possible to confidently predict the subjective judgement of these loudspeakers. New input features have been developed and combined with existing sound metrics and raw measurements to provide the inputs. The program utilises cascaded ensemble methods, and is simplified to ensure a level of complexity appropriate to the data. Automatic anomaly detection removes potentially misleading data before re-calculating the most suitable decision boundary, based on those units most important to the groupings. The final performance based on separate test data shows a marked improvement relative to suitable alternative methods, and a drastic improvement compared to analysis of any one measurement alone. The output is a classification which correlates highly to subjective judgements, based on previously agreed criteria.

7 REFERENCES

1. P. Newell and K. Holland, Loudspeakers: for Music Recording and Reproduction, Focal Press. (2007).
2. H. Fastl, 'Psychoacoustic Basis of Sound Quality Evaluation and Sound Engineering', ICSV13, Vienna, Austria. (2006).
3. M.N. Valenzuela, 'Untersuchungen und Berechnungsverfahren zur Klangqualität von Klaviertönen', Herbert Utz Verlag Wissenschaft, Munich. (1998).
4. H. Fastl and E. Zwicker, Psychoacoustics - Facts and Models, 3rd ed Springer. (2007).
5. H. Staffeldt, 'Correlation between Subjective and Objective Data for Quality Loudspeakers', JAES Volume 22 Issue 6 pp. 402-415. (1974).
6. W. Klippel, 'Multidimensional Relationship between Subjective Listening Impression and Objective Loudspeaker Parameters', Acustica, Vol. 70, pp. 45-54. (1990).
7. S. Olive, 'A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part I - Listening Test Results', presented at the 116th AES Convention, Berlin, Germany, preprint 6113. (2004)
8. S. Olive, 'A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II - Development of the Model', presented at the 117th AES Convention, San Francisco, USA, preprint 6190. (2004).
9. F. Toole, 'Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 2', JAES Volume 34 Issue 5 pp. 323-348. (1986).
10. F. Toole, Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms, 1st ed Focal Press. (2008).
11. G. Peters, 'A large set of Audio Features for Sound Description' (white paper), Ircam. (2004).
12. J.F. Hair Jr et al, Multivariate Data Analysis, 7th ed Pearson. (2008).
13. N. Christianini and J.C. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press. (2000).
14. P. Viola and M. Jones, 'Robust Real-time Object Detection', International Journal of Computer Vision. (2001)