ROBUSTNESS AND INDEPENDENCE OF VOICE TIMBRE FEATURES UNDER LIVE PERFORMANCE ACOUSTIC DEGRADATIONS

Dan Stowell and Mark D. Plumbley

Centre for Digital Music,
Dept. of Electronic Engineering,
Queen Mary, University of London, UK
dan.stowell@elec.qmul.ac.uk

ABSTRACT

Live performance situations can lead to degradations in the vocal signal from a typical microphone, such as ambient noise or echoes due to feedback. We investigate the robustness of continuous-valued timbre features measured on vocal signals (speech, singing, beatboxing) under simulated degradations. We also consider non-parametric dependencies between features, using information theoretic measures and a feature-selection algorithm. We discuss how robustness and independence issues reflect on the choice of acoustic features for use in constructing a continuous-valued vocal timbre space. While some measures (notably spectral crest factors) emerge as good candidates for such a task, others are poor, and some features such as ZCR exhibit an interaction with the type of voice signal being analysed.

1. INTRODUCTION

Real-time analysis of the timbre of a voice signal may provide useful information for musical control, effects, or interactive systems. Many acoustic features are available for timbre analysis (see e.g. [1]). In this study we investigate various features and their suitability for analysing the signals from vocal performances, where suitability is judged in two ways: robustness to signal degradations that may occur in live performances; and the amount of independent information they contribute, determined using information-theoretic methods.

"Timbre" is relatively ambiguous as a concept: some timbral analyses are based purely on the harmonic strengths of pitched sounds [2] while some are based on pitch-agnostic methods such as MFCCs [3]. Evidence from perceptual studies tells us that timbre perception is multidimensional and probably non-linear [4][5].

Our concern in this study is to evaluate continuous-valued and pitch-agnostic instantaneous timbre features, which may be useful in constructing a multidimensional "timbre space". Many others have evaluated timbre features in the context of classification tasks, such as audio genre classification [6], beat detection [7] or speech recognition [8]. In this study we do not perform any classification, although a continuous timbre space could in principle be used as input to a classifier.

We first describe the voice data used and the features investigated, then present two experiments: one on the robustness of the features to acoustic degradations, and one to investigate information-theoretic dependencies between features.

2. DATA PREPARATION

For our experiments we prepared three datasets representing three types of performing voice: singing, speech, and beatboxing. Participants were aged 18–40 and with varying levels of musical training. For the singing and speech datasets we recorded 5 male and 3 female participants; for the beatbox datasets we recorded 4 male participants (the beatboxing community is predominantly male). All recordings were made in an acoustically-treated studio, using a Shure SM58 microphone and Focusrite Red 1 preamp, recorded at 44.1 kHz and 32-bit resolution. Each recording was amplitude-normalised and long pauses were removed.

Feature analysis was then performed in SuperCollider 3.2 [9], segmenting sounds into 1024-sample frames (with 50% overlap between frames) and using a Hann window for FFT-based features. Low-power frames (silences) were removed from the analysis.

The total number of audio frames in each dataset was then approximately:

- Singing: 878,000 frames
- Speech: 987,000 frames
- Beatboxing: 454,000 frames

2.1. Features investigated

From the three datasets we derived the following features for each frame (for definitions see [1]):

- Eight MFCCs, derived from 42 Mel-spaced filters
- Spectral centroid (power-weighted mean frequency)
- Spectral spread (power-weighted standard deviation)
- Spectral crest factor (SCF)
- Spectral crest factor in four log-spaced subbands (50–400, 400–800, 800–1600, and 1600–3200 Hz)
- Spectral distribution percentiles: 25%, 50%, 90%, 95% (the latter two are often described as "spectral rolloff" measures)
- High-frequency content (HFC)
- Zero-crossing rate (ZCR)
- Spectral flatness
- Spectral flux

In preliminary tests we also investigated some additional features (using a smaller voice dataset): energy ratio in log-spaced and in ERB-spaced subbands, SCF in higher-frequency subbands, some other percentiles of the power spectrum, and HFC normalised against frame power. These features gave very poor robustness performance compared against other features, so we did not include them in the main experiments.

In our preliminary tests we also visualised the distribution of feature values and found by inspection that many of them were not normally-distributed. Therefore all analyses we use in the following experiments are based on non-parametric statistics.

3. EXPERIMENT 1: ROBUSTNESS TO DEGRADATIONS

To investigate the robustness of timbre features to signal degradation, we applied the following degradations to the datasets, each at 4 levels of effect:

- Additive white noise
- Additive crowd noise
  (a "club crowd" recording from a commercial sample CD)
- Additive music noise
  ("Come Back Clammy Lammy" by The Cardiacs)
- Clipping distortion
- Delay with no feedback
- Delay with feedback
- Reverberation (FreeVerb)

For each dataset this therefore created 7 x 4 = 28 degraded versions.

We then measured, for each audio frame in each degraded recording, the absolute percentage deviation of the timbre features from their "clean" values (taken from the original recording).

To gain an overview of the relative performance of each timbre feature, we performed some statistics across the whole 28 degradations per dataset:

- Kendall's W test [10] to determine the extent to which some features' deviation was consistently better or worse than others'. This statistic looks just at the relative ranking of features' deviation within a frame, ignoring the magnitudes of differences between deviations.
- For each pair of features, the Wilcoxon Signed Rank test [11, section 15.4] to determine whether one feature performs better than another in terms of deviation, and if so how strong the difference between the two is. This is intended to "drill down" beyond Kendall's W test to look at the magnitude of differences in performance.

For practical reasons the data frames were subsampled by a factor of 10 before calculating these statistics.

Note that these statistics amalgamate the performance of features across different effect levels, which may mask an interaction between robustness and effect level: a given feature might perform extremely well under mild degradation but fail catastrophically under high degradation. We therefore plotted graphs (not shown) of the deviation values under the different settings. By visual inspection, the dominant trend was that features which performed best at mild effect level also tended to perform best at strong effect level.

3.1. Results of experiment 1

3.1.1. Overall tests

Tables 1, 2 and 3 show (for each dataset) the median deviation of the features, as well as the median ranking of the features within each frame (ranked by deviation within that frame). For the latter (the median rank), Kendall's W test indicated statistically significant rankings ($p < 0.0001$) in each of the three datasets. Each table is sorted according to median deviation.

Feature	Med. deviation (%)	Med. rank (W=0.274)
crst1	2.17	5
25%ile	3.64	7
crst2	4.63	7
ZCR	5.18	9
mfcc1	6.62	9
95%ile	6.73	9
spread	7.01	8
crest	7.4	9
50%ile	7.71	10
crst3	10.9	10
90%ile	11.3	11
centroid	11.3	11
mfcc3	14.6	12
crst4	14.7	11
mfcc5	16.9	13
mfcc8	19.3	14
mfcc7	19.8	14
flatness	21.2	15
mfcc4	28.1	16
mfcc2	31.3	17
flux	36.6	17
mfcc6	40.9	18
HFC	6.33e+05	23

Table 1: Median deviation of each feature over all degradations; and the median rank obtained for each feature, when ranked according to lowest deviation for each audio frame. (In both columns smaller values are better.) Singing dataset.

The tables indicate similarities across the datasets but also some differences. In general it seems that the HFC, flux, flatness and some even-numbered MFCCs are the worst-performing, showing the highest typical deviation and the worst typical ranking.[1] The best-performing feature on all three datasets was *crst1* (spectral crest in the lowest subband), typically followed by odd-numbered MFCCs and others of the spectral crest features.

However, the ranking of some features shows notable differences across the datasets. The ranking of ZCR, 25-percentile and 95-percentile seems relatively high for the singing dataset, but then is lower for the speech dataset and lower still for the beatboxing dataset. We will discuss this further in section 5.

3.1.2. Pairwise tests

The results of applying the Wilcoxon Signed-Rank test to pairs of features are shown in Tables 4, 5, and 6, formatted so that a plus (+) sign indicates that the column feature has outperformed the row feature with a z-score significant at the $p < 0.0001$ level, and

[1] Even-numbered MFCCs differ from odd-numbered MFCCs in the effect of the Discrete Cosine Transform: for even-numbered MFCCs the coefficient includes a positive contribution from the highest Mel-frequency bins, while for odd-numbered MFCCs the contribution is negative.

wSR-Z	25%ile	50%ile	90%ile	95%ile	centroid	spread	flatness	ZCR	HFC	flux	crest	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8	crst1	crst2	crst3	crst4
25%ile		−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
50%ile	+		−	+	−	+	−	+	−	−	+	+	−	−	−	−	−	−	−	+	+	−	−
90%ile	+	+		+	+	+	−	+	−	−	+	+	−	−	−	−	−	−	−	+	+	+	−
95%ile	+	−	−		−	−	−	+	−	−	−	+	−	−	−	−	−	−	−	+	+	−	−
centroid	+	+	−	+		+	−	+	−	−	+	+	−	−	−	−	−	−	−	+	+	+	−
spread	+	−	−	+	−		−	+	−	−	−	+	−	−	−	−	−	−	−	+	+	−	−
flatness	+	+	+	+	+	+		+	−	−	+	+	−	+	−	+	−	+	+	+	+	+	+
ZCR	+	−	−	−	−	−	−		−	−	−	−	−	−	−	−	−	−	−	+	+	−	−
HFC	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+
flux	+	+	+	+	+	+	+	+	−		+	+	+	+	+	+	−	+	+	+	+	+	+
crest	+	−	−	+	−	+	−	+	−	−		+	−	−	−	−	−	−	−	+	+	−	−
mfcc1	+	−	−	−	−	−	−	+	−	−	−		−	−	−	−	−	−	−	+	+	−	−
mfcc2	+	+	+	+	+	+	+	+	−	−	+	+		+	+	+	−	+	+	+	+	+	+
mfcc3	+	+	+	+	+	+	−	+	−	−	+	+	−		−	−	−	−	−	+	+	+	−
mfcc4	+	+	+	+	+	+	+	+	−	−	+	+	−	+		+	−	+	+	+	+	+	+
mfcc5	+	+	+	+	+	+	−	+	−	−	+	+	−	+	−		−	−	−	+	+	+	+
mfcc6	+	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+		+	+	+	+	+	+
mfcc7	+	+	+	+	+	+	−	+	−	−	+	+	−	+	−	+	−		+	+	+	+	+
mfcc8	+	+	+	+	+	+	−	+	−	−	+	+	−	+	−	+	−	−		+	+	+	+
crst1	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−		−	−	−
crst2	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+		−	−
crst3	+	+	−	+	−	+	−	+	−	−	+	+	−	−	−	−	−	−	−	+	+		−
crst4	+	+	+	+	+	+	−	+	−	−	+	+	−	+	−	−	−	−	−	+	+	+	

Table 4: z-scores for Wilcoxon Signed Rank test, between pairs of features. A diagonal line indicates that no difference between feature distributions is proven ($p > 0.0001$) for that pair. Otherwise, a plus sign indicates the column feature deviates less, and a minus sign indicates it deviates more, than the row feature. Singing dataset.

wSR-Z	25%ile	50%ile	90%ile	95%ile	centroid	spread	flatness	ZCR	HFC	flux	crest	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8	crst1	crst2	crst3	crst4
25%ile		−	−	−	−	−	−	−	−	−	−	+	−	/	−	−	−	−	−	+	+	−	−
50%ile	+		−	−	/	+	−	−	−	−	+	+	−	−	−	−	−	−	−	+	+	−	−
90%ile	+	+		+	+	+	−	+	−	−	+	+	−	+	−	+	−	+	−	+	+	+	+
95%ile	+	+	−			+	−	+	−	−	+	+	−	+	−	+	−	+	−	+	+	+	−
centroid	+	/	−			+	−	+	−	−	+	+	−	+	−	+	−	+	−	+	+	+	+
spread	+	−	−	−	−		−	−	−	−	−	+	−	−	−	−	−	−	−	+	+	−	−
flatness	+	+	+	+	+	+		+	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+
ZCR	+	+	−	−	−	+	−		−	−	+	+	−	−	−	+	−	−	−	+	+	+	−
HFC	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+
flux	+	+	+	+	+	+	+	+	−		+	+	+	+	+	+	+	+	+	+	+	+	+
crest	+	−	−	−	−	+	−	−	−	−			−	−	−	−	−	−	−	+	+	−	−
mfcc1	−	−	−	−	−	−	−	−	−	−			−	−	−	−	−	−	−	+	−	−	−
mfcc2	+	+	+	+	+	+	−	+	−	−	+	+		+	+	+	+	+	+	+	+	+	+
mfcc3	/	+	−	−	−	+	−	+	−	−	+	+	−		−	+	−	+	−	+	+	+	−
mfcc4	+	+	+	+	+	+	−	+	−	−	+	+	−	+		+	−	+	−	+	+	+	+
mfcc5	+	+	−	−	−	+	−	−	−	−	+	+	−	−	−		−	−	−	+	+	−	−
mfcc6	+	+	+	+	+	+	−	+	−	−	+	+	−	+	+	+		+	+	+	+	+	+
mfcc7	+	+	−	−	−	+	−	+	−	−	+	+	−	−	−	+	−		−	+	+	+	−
mfcc8	+	+	+	+	+	+	−	+	−	−	+	+	−	+	+	+	−	+		+	+	+	+
crst1	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−		−	−	−
crst2	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	+		−	−
crst3	+	+	−	−	−	+	−	−	−	−	+	+	−	−	−	+	−	−	−	+	+		−
crst4	+	+	−	+	−	+	−	+	−	−	+	+	−	+	−	+	−	+	−	+	+	+	

Table 5: z-scores for Wilcoxon Signed Rank test, between pairs of features. A diagonal line indicates that no difference between feature distributions is proven ($p > 0.0001$) for that pair. Otherwise, a plus sign indicates the column feature deviates less, and a minus sign indicates it deviates more, than the row feature. Speech dataset.

wSR-Z	25%ile	50%ile	90%ile	95%ile	centroid	spread	flatness	ZCR	HFC	flux	crest	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8	crst1	crst2	crst3	crst4
25%ile		−	−	−	−	+	−	−	−	−	+	+	−	+	+	+	+	+	+	+	−	−	−
50%ile	+		−	−	−	+	−	−	−	−	+	+	−	+	+	+	+	+	+	+	+	+	−
90%ile	+	+		+	+	+	−	−	−	−	+	+	−	+	+	+	+	+	+	+	+	+	+
95%ile	+	+	−		−	+	−	−	−	−	+	+	−	+	+	+	+	+	+	+	+	+	−
centroid	+	+	−	+		+	−	−	−	−	+	+	−	+	+	+	+	+	+	+	+	+	+
spread	−	−	−	−	−		−	−	−	−	+	+	−	+	−	+	−	+	+	+	−	−	/
flatness	+	+	+	+	+	+		+	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+
ZCR	+	+	+	+	+	+	−		−	−	+	+	−	+	+	+	+	+	+	+	+	+	+
HFC	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+
flux	+	+	+	+	+	+	+	+	−		+	+	+	+	+	+	+	+	+	+	+	+	+
crest	−	−	−	−	−	−	−	−	−	−		+	−	+	−	+	−	+	−	+	−	−	−
mfcc1	−	−	−	−	−	−	−	−	−	−	−		−	−	−	+	−	+	−	+	−	−	/
mfcc2	+	+	+	+	+	+	−	+	−	−	+	+		+	+	+	+	+	+	+	+	+	+
mfcc3	−	−	−	−	−	−	−	−	−	−	−	+	−		−	+	−	+	−	+	−	−	−
mfcc4	−	−	−	−	−	+	−	−	−	−	+	+	−	+		+	+	+	+	+	−	−	−
mfcc5	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−		−	−	−	+	−	−	−
mfcc6	−	−	−	−	−	+	−	−	−	−	+	+	−	+	−	+		+	+	+	−	−	−
mfcc7	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−		−	+	−	−	−
mfcc8	−	−	−	−	−	−	−	−	−	−	+	+	−	+	−	+	−	+		+	−	−	−
crst1	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−		−	−	−
crst2	+	−	−	−	−	+	−	−	−	−	+	+	−	+	+	+	+	+	+	+		−	−
crst3	+	−	−	−	−	+	−	−	−	−	+	+	−	+	+	+	+	+	+	+	+		−
crst4	+	+	−	+	−	/	−	−	−	−	+	/	−	+	+	+	+	+	+	+	+	+	

Table 6: z-scores for Wilcoxon Signed Rank test, between pairs of features. A diagonal line indicates that no difference between feature distributions is proven ($p > 0.0001$) for that pair. Otherwise, a plus sign indicates the column feature deviates less, and a minus sign indicates it deviates more, than the row feature. Beatboxing dataset.

a minus ($-$) sign indicates the reverse, that the row feature has outperformed the column feature. General tendencies confirm what is seen in the median-rank tables, for example the very weak performance of the HFC and flux features, which rarely outperform any other feature on any dataset. The lowest-subband SCF (crst1) performs very strongly on all datasets, and some of the other SCF features show a tendency to outperform other features in terms of robustness. Some MFCCs also perform strongly.

However, not all features show the same performance across datasets: the ZCR seems relatively robust on the singing dataset, outperforming many other features, while for the speech dataset its performance is mixed and for the beatboxing dataset it is worse-performing than most other features. This confirms what was seen in the median-ranking data.

Another notable difference across the datasets is the relative performance of the MFCCs. On the beatboxing dataset, the MFCCs show strong robustness (especially the odd-numbered coefficients), outperforming most other features in the pairwise comparisons. However the singing dataset shows a different result: the MFCCs are outperformed by the percentile features, by centroid and spread, and by most of the SCF features. Results for the speech dataset lie somewhere between the two: the even-numbered MFCCs perform worse than most other features, while the odd-numbered MFCCs perform better. We will return to these differences in section 5.

4. EXPERIMENT 2: FEATURE INTERDEPENDENCE

Using the same datasets and features, we also investigated the extent of interdependence and redundancy between timbre features.

The non-parametric distribution of timbre features suggests that parametric analyses (e.g. correlation) may be problematic. Instead, we applied information-theoretic measures calculated using model-free methods. Since we are interested in the application of timbre features in live performance situations, we performed our information-theoretic analyses across the expanded datasets consisting of the “clean” recording plus 28 degraded recordings per dataset.

Firstly, to investigate relationships between pairs of features we determined the *mutual information* between each pair of features. The mutual information is given as

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\, p(y)} \right) \tag{1}$$

where $p(x,y)$ is the joint probability distribution of X and Y and $p(x)$ is the marginal probability distribution of X. We calculated this using an adaptive partitioning method with 16 bins per feature, giving a range of 0 to 4 bits for the mutual information value.

Secondly, in order to provide an information-theoretic ranking of the contributions of the timbre features, we performed a “greedy rejection” feature selection algorithm on each dataset, as follows. Starting with the full set of features, we evaluate the conditional entropy of each feature (the entropy conditional on all the other features):

$$H(X|A) = H(X,A) - H(A) \tag{2}$$

where X indicates a feature, A indicates all features other than X and $H(.)$ is the entropy. The conditional entropy in this context indicates the uncertainty of a feature’s value given the values of the other features, and therefore the amount of “extra information” that the feature adds to the ensemble. For high-dimensional entropy

Feature	Med. deviation (%)	Med. rank (W=0.102)
crst1	7.24	7
mfcc1	8.9	9
crst2	9.32	8
25%ile	10.2	10
spread	11.2	9
crest	11.3	9
50%ile	12.1	11
mfcc5	14.8	11
crst3	15.4	10
ZCR	16.2	12
mfcc7	16.7	12
mfcc3	17	12
95%ile	17.2	12
centroid	17.3	12
crst4	17.3	11
90%ile	19	13
mfcc4	20	13
mfcc8	24.1	15
mfcc2	29.9	16
mfcc6	29.9	16
flatness	30.8	17
flux	33	16
HFC	47.7	19

Table 2: Median deviation of each feature over all degradations; and the median rank obtained for each feature, when ranked according to lowest deviation for each audio frame. (In both columns smaller values are better.) Speech dataset.

Feature	Med. deviation (%)	Med. rank (W=0.0887)
crst1	9.74	8
mfcc5	10.1	9
mfcc7	10.2	9
mfcc1	10.9	10
mfcc3	11.7	10
crest	12.6	11
mfcc8	12.9	11
spread	13.2	10
mfcc6	14	12
mfcc4	14.1	11
25%ile	15.4	13
crst2	15.5	11
crst3	16.7	12
50%ile	18.3	14
95%ile	18.4	12
crst4	18.9	13
centroid	20.7	13
90%ile	23.4	14
ZCR	23.5	14
mfcc2	24.5	15
flatness	34	17
flux	41.9	18
HFC	46.7	20

Table 3: Median deviation of each feature over all degradations; and the median rank obtained for each feature, when ranked according to lowest deviation for each audio frame. (In both columns smaller values are better.) Beatboxing dataset.

estimation we used an adaptive partitioning method (related to that of [12]) which estimates the differential entropy. Hence entropy values can be negative, and depend on the scaling of the feature ranges. In this test we therefore normalised feature values to a range of 0–1.

We identify the feature(s) whose conditional entropy is the lowest, and remove it from consideration. We then repeat the procedure using the now-smaller set of features, and continue the process until only one feature remains. (Note that conditional entropy of a feature depends on the context, i.e. which features are included in A. It must be recalculated at each step since discarding one feature changes the context for the remainders.)

A practical limitation in our entropy estimation algorithm meant that it could not analyse the full 24-dimensional space without a much larger volume of data. We therefore excluded four of the least-robust features (as determined in Experiment 1): power, flatness, HFC, flux and 2nd MFCC. We subsampled the data by a factor of 3.

4.1. Results of experiment 2

4.1.1. Pairwise mutual information

The mutual information between pairs of features showed the same general tendencies in all three datasets, so we include here only one table, the results from the speech dataset (Table 7). Note that the values in Table 7 are expressed as percentages for ease of reading.

Five features seem to have a high interdependence, with at least 1.1 bits of information shared between any pair: centroid, spread, flatness, 95-percentile and first MFCC. This overlap is interesting because the features represent three alternative approaches to spectral analysis: parametric statistics (centroid, spread), nonparametric statistics (percentiles, flatness) and perceptually-inspired calculations (MFCC). The interaction of the first MFCC is notable since all the other MFCCs show extremely low amounts of mutual information with any other features. The first MFCC essentially encodes the low-versus-high balance of energy in the Mel-scaled spectrum; and so it seems likely that the interaction of these five features is because each of them has a connection to the low-versus-high balance of spectral energy when analysing voice signals.

Some other pairs of features share information, though to a lesser extent: all pairs of percentile features share some information, and both HFC and flux indicate a degree of interaction with the signal power. The HFC and flux calculations have similarities to the power calculation (HFC is essentially a weighted power calculation) so their interaction is understandable.

MFCCs 2–8 and the SCF measures are generally quite independent, not showing large interaction with any other features. (The full-band crest measure exhibits some interaction with crst1, suggesting that the full-band crest measure is most influenced by the content of that lowest band.)

$I(X;Y)$	power	25%ile	50%ile	90%ile	95%ile	centroid	spread	flatness	ZCR	HFC	flux	crest	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8	crst1	crst2	crst3
crst4	2.1	3.7	4.4	4.5	4.4	4.6	4.2	4.7	2.8	4.2	2.2	3.4	4.2	0.6	1.1	1	0.5	1.6	0.9	2.2	0.6	1.2	1.9
crst3	1.5	2.7	3.3	3.2	2.9	3.3	2.6	3.2	2.3	2.9	1.8	2.6	3.1	0.4	0.5	0.7	1.3	2.3	0.3	0.6	0.7	1.1	
crst2	1.4	2.7	4	1.9	1.6	2.3	1.4	1.9	2.4	2.4	1.9	4.6	2.1	0.6	1	1.7	1	1.4	1.7	2.2	0.6		
crst1	0.8	1.2	1.6	0.8	0.8	1	0.7	1	1.3	1	0.6	11	1.1	0.6	0.4	0.3	0.3	1.2	0.5	0.6			
mfcc8	1.8	3.3	3.9	3.1	3.1	3.4	2.4	3.2	1.8	4	2.1	3.3	2.2	1	1	0.7	0.4	2.8	2.8				
mfcc7	0.9	2	2.4	2.2	2.3	2.2	2	2.5	1.2	2.5	1.1	2.1	2.3	0.7	0.8	1	1.3	1.8					
mfcc6	1.9	5.7	5.9	5.5	5.1	5.2	4.9	6	3.3	5	1.8	3.7	4.4	0.9	1.2	1.2	1.8						
mfcc5	1.4	4	5	3.4	3.5	2.8	3.2	3.3	2.3	2.6	1.4	2.1	2.3	0.8	3.6	3.5							
mfcc4	1.7	5.3	6.7	2.7	2.3	2.8	2.1	2.8	2.9	3.7	2.4	2.9	1.6	0.7	2.7								
mfcc3	2.6	5.6	8.4	7.1	6.1	5.4	5.5	5.5	3.2	4.7	2.4	3.8	6.2	1.7									
mfcc2	2.4	9.1	12	10	6.6	9.8	4.2	4.4	4.6	7	3	3.7	3.8										
mfcc1	5.2	16	20	**35**	**37**	**35**	**35**	**31**	11	17	5.9	10											
crest	5.7	16	17	9.6	8.6	12	7.6	9.3	10	11	6.3												
flux	19	7.8	9	6.6	6.4	7.9	5.4	6.9	4.7	15													
HFC	14	18	24	20	18	25	15	16	11														
ZCR	3.4	16	19	12	9.8	14	7.9	8.9															
flatness	6.8	12	16	**30**	**36**	**31**	**48**																
spread	5.1	9.8	13	**31**	**42**	**27**																	
centroid	6.6	19	**30**	**48**	**36**																		
95%ile	5.3	12	17	**44**																			
90%ile	5.6	14	21																				
50%ile	7.2	**32**																					
25%ile	6.7																						

Table 7: Mutual Information between features, for the speech dataset. Values are expressed as a percentage of the maximum possible value of 4 bits. Values greater than 25% (i.e. 1 bit) are given in bold.

4.1.2. *Feature selection*

Results of the entropy-based feature-selection are given in Tables 8 (for singing), 9 (for speech) and 10 (for beatboxing).

The three tables show some common trends. The most-favoured features are spectral crest measurements (full-band or subband), MFCCs, and the 95-percentile.

The ZCR is the only time-domain feature included here, and so one might have expected it to provide information in some sense different from the spectral features. However it is rejected quite early in all three trials, indicating this is not the case. We suggest that this may be because the fluctuations in ZCR are affected by vowel/consonant distinctions in almost the same way as measures like spectral rolloff or centroid: vowels tend to have a low value while consonants (especially fricatives) push it to a high value.

The low percentiles were rejected very quickly in our experiments, suggesting that they do not carry much information at all that cannot be derived from others of our features used. This provides an interesting contrast against the 95-percentile, which performs quite well (in the speech dataset, very well).

The poor performance of the spectral centroid measure is notable given its relatively common usage in topics such as timbre analysis and music retrieval. Compare it against those features with which it has a high mutual information (as discussed above): our feature selection consistently ranks it below the spectral spread, 95-percentile and first MFCC.

Across the three datasets very little difference is evident. The feature with the largest change in ranking is the 95-percentile – ranked much higher for the speech dataset than the others – but all other features exhibit quite a consistent performance in terms of this feature-selection experiment.

As discussed, we were unable to run the feature selection on the full 24 features, and excluded the least noise-robust from the main feature selection test reported here. In order to check whether or not we had unjustly excluded features which may have carried much information, we ran a second feature selection test, returning the excluded features and instead excluding a different subset – namely, the poorly-ranking features as shown in Tables 8, 9 and 10. In this test (data not shown) we found that most of the excluded features (power, flatness, HFC, flux) ranked very badly, consistently among the first to be rejected. The 2nd MFCC achieved a middling rank, alongside other MFCCs.

5. DISCUSSION

Our two experiments each compare timbre features but using very different criteria. Some features perform well according to both sets of criteria: in particular the (full-band or subband) spectral crest features, which we find to be quite noise-robust as well as information-bearing. This suggests that they can be recommended generally for analysis of voice signals. Odd-numbered MFCC coefficients also performed strongly in both tests.

The strong performance of the spectral crest features is in concordance with experiments in music similarity [13], audio retrieval [14], and speaker recognition [15], which have found them to be useful when included alongside other features such as MFCCs. Our feature-selection experiment confirms this for the case of the performing voice, and also reinforces the notion that spectral crest features may specifically be *complementary* to MFCCs, because in our feature-selection experiment both the MFCC and the SCF feature-sets performed strongly – implying that neither is strongly predictable from the other.

Some features perform badly in both experiments. The spectral flatness, spectral flux, and HFC measures are all relatively well-known features, yet when applied to our voice datasets they seem to be highly susceptible to acoustic degradations, as well as

Rank	Feature	H(Feature|Remaining)
1	crst2	—
2	crst3	-0.871
3	crest	-0.854
4	mfcc6	-0.759
5	mfcc8	-0.636
6	mfcc3	-0.391
7	crst1	-0.166
8	mfcc7	0.0196
9	95%ile	0.0425
10	mfcc4	0.0127
11	mfcc5	0.365
12	mfcc1	0.0722
13	spread	0.0962
14	90%ile	0.403
15	crst4	0.0392
16	centroid	0.13
17	ZCR	-0.323
18	50%ile	-0.742
19	25%ile	-1.35

Table 8: Feature selection by greedy rejection using conditional entropy, for singing dataset.

Rank	Feature	H(Feature|Remaining)
1	crst2	—
2	95%ile	-0.884
3	crst1	-0.814
4	crst3	-0.891
5	mfcc8	-0.773
6	mfcc3	-0.527
7	mfcc7	-0.435
8	mfcc6	-0.00864
9	mfcc4	0.0394
10	mfcc5	-0.188
11	crest	0.217
12	mfcc1	0.196
13	spread	0.175
14	90%ile	0.313
15	crst4	-0.0589
16	centroid	0.0974
17	ZCR	-0.254
18	50%ile	-0.845
19	25%ile	-1.23

Table 9: Feature selection by greedy rejection using conditional entropy, for speech dataset.

carrying a lower amount of unique information than other features we investigated.

For some features the experimental results point in different directions. Even-numbered MFCC coefficients were ranked relatively highly in the entropy-based feature selection, but ranked poorly in terms of robustness to degradations. Conversely, spectral spread ranked quite highly for robustness but poorly for information content, as did the 25-percentile and 50-percentile.

We were interested to see the relative performance of the 90- and 95-percentile, since both have been used as "spectral rolloff" measures (with 95-percentile the more common). The 95-percentile consistently outranked the 90-percentile in both of our experiments, confirming that is is preferable, for solo voice data at least.

The above discussion has focussed on findings that emerge consistently across the three datasets. However, Experiment 1 revealed differences across datasets, in terms of the robustness of some features. For the singing dataset the 25- and 95-percentiles and the ZCR give good robustness performance, and the MFCCs are less robust than many other features; while for the beatboxing dataset the reverse is true; and for speech the results lie somewhere between the two extremes. We suggest that vowel/consonant differences may be the cause of this: compared against speech, singing contains a larger proportion of vowel phonation, whereas beatboxing involves a much smaller proportion [16]. Vowels and consonants are acoustically very different classes of sound, producing spectral structures very different in gross and fine detail [17]. Therefore it seems likely that some features are more robust when analysing vowels than consonants.

Our results suggest which features may be more or less useful in constructing a timbre space for vocal signals. However, there are some questions which remain unanswered. The effect of signal degradation on separate timbre features is likely to exhibit strong interactions, so the deviations of individual features don't allow us to predict directly the deviation within a multidimensional timbre space made from those features. The issue becomes further complicated if data-reduction techniques such as Principal Component Analysis, Independent Component Analysis or Self-Organising Maps are applied. This is a topic for future study; as is the question of how easily a performer can exert deliberate control over such timbre dimensions and spaces.

6. CONCLUSIONS

We have evaluated timbre features for use with performing voice signals, according to two criteria: their robustness to acoustic degradations of the signal – such as might occur in a performance situation – and their statistical independence from other features (and therefore the amount of "extra information" they provide). The strongest-performing features were the (full-band or subband) spectral crest factors (SCFs) and the odd-numbered MFCCs, which generally exhibited good robustness as well as being informative in the information-theoretic sense. The 95-percentile also performed quite strongly, and is to be recommended as a measure of "spectral rolloff".

Some features performed poorly in both evaluations, including spectral flatness, spectral flux, and HFC. In the context of our voice datasets they showed poor robustness as well as a degree of informational overlap with other features. Although these features are relatively well-known in music signal analysis, we suggest that they should be avoided for analysis of monophonic voice signals.

Similarly, the spectral centroid feature is relatively common in music signal analysis, yet its mediocre performance on our voice datasets suggests that it may not be especially useful for solo voice analysis. It is generally outperformed by features such as 95-percentile and first MFCC, with which it has quite a high informational overlap.

The performance of some features (e.g. ZCR) showed an interaction with signal type, possibly due to the differing nature of

Rank	Feature	H(Feature|Remaining)
1	crst1	—
2	mfcc1	-0.823
3	crst2	-0.796
4	mfcc5	-0.909
5	mfcc7	-0.787
6	mfcc3	-0.571
7	mfcc8	-0.195
8	mfcc4	0.0262
9	mfcc6	-0.101
10	crest	0.381
11	spread	0.494
12	crst3	0.264
13	95%ile	0.207
14	crst4	0.318
15	90%ile	-0.398
16	centroid	-0.053
17	ZCR	-0.475
18	50%ile	-1.1
19	25%ile	-1.57

Table 10: Feature selection by greedy rejection using conditional entropy, for beatboxing dataset.

vowel- and consonant-type sounds.

More broadly, we have shown that information-theoretic measures can be useful in answering questions about the interactions between acoustic features, whose non-parametric distributions may lead to problems for more traditional measures of association.

7. ACKNOWLEDGMENTS

We would like to thank Samer Abdallah for useful discussions on information-theoretic methods.

8. REFERENCES

[1] G. Peeters, “A large set of audio features for sound description,” Tech. Rep., IRCAM, 2004.

[2] H. F. Pollard and E. V. Jansson, “A tristimulus method for the specification of musical timbre,” *Acustica*, vol. 51, pp. 162–171, 1982.

[3] J.-J. Aucouturier and F. Pachet, “Improving timbre similarity: How high is the sky?,” *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, 2004.

[4] J. M. Grey and J. W. Gordon, “Perceptual effects of spectral modifications on musical timbres,” *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1493–1500, 1978.

[5] A. Burgoyne and S. McAdams, “Non-linear scaling techniques for uncovering the perceptual dimensions of timbre,” in *Proceedings of the International Computer Music Conference (ICMC’07)*, Copenhagen, Denmark, August 2007, vol. 1, pp. 73–76.

[6] M. F. McKinney and J. Breebaart, “Features for audio and music classification,” *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, pp. 151–158, 2003.

[7] F. Gouyon, S. Dixon, and G. Widmer, “Evaluating low-level features for beat classification and tracking,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP’07)*, 2007.

[8] B. Kotnik, D. Vlaj, and B. Horvat, “Efficient noise robust feature extraction algorithms for Distributed Speech Recognition (DSR) systems,” *International Journal of Speech Technology*, vol. 6, pp. 205–219, 2003.

[9] J. McCartney, “Rethinking the computer music language: SuperCollider,” *Computer Music Journal*, vol. 26, no. 4, pp. 61–68, 2002.

[10] M. G. Kendall and B. B. Smith, “The problem of m rankings,” *The Annals of Mathematical Statistics*, vol. 10, no. 3, pp. 275–287, Sep 1939.

[11] W. Mendenhall, D. D. Wackerly, and R. L. Scheaffer, *Mathematical statistics with applications*, PWS-Kent, fourth edition, 1989.

[12] E. G. Learned-Miller, “A new class of entropy estimators for multi-dimensional densities,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP’03)*, April 2003, vol. 3, pp. 297–300.

[13] J. Herre, E. Allamanche, and C. Erie, “How similar do songs sound? towards modeling human perception of musical similarity,” in *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-2003)*, October 2003, pp. 83–86.

[14] J. Herre, E. Allamanche, and O. Hellmuth, “Robust matching of audio signals using spectral flatness features,” in *Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA-2001)*, 2001, pp. 127–130.

[15] D. Hosseinzadeh and S. Krishnan, “On the use of complementary spectral features for speaker recognition,” *EURASIP Journal on Advances in Signal Processing*, vol. 2008 (Article ID 258184, 10 pages), 2008.

[16] D. Stowell and M. D. Plumbley, “Characteristics of the beatboxing vocal style,” Tech. Rep. C4DM-TR-08-01, Dept. of Electronic Engineering, Queen Mary, University of London, 2008.

[17] D. B. Fry, *The Physics of Speech*, Cambridge Textbooks in Linguistics. Cambridge University Press, 1996.