ISCA Archive
http://www.isca-speech.org/archive

ITRW on
Speech and Emotion
Newcastle, Northern Ireland, UK
September 5-7, 2000

APPROACHING AUTOMATIC RECOGNITION OF EMOTION FROM VOICE: A ROUGH BENCHMARK

Sinéad McGilloway [1], Roddy Cowie[2], Ellen Douglas-Cowie[2],

Stan Gielen[3], Machiel Westerdijk[3] and Sybert Stroeve[3]

[1] National University of Ireland, Maynooth
[2] Queen's University of Belfast, Northern Ireland
[3] University of Nijmegen

ABSTRACT

Automatic recognition of a speaker's emotions is a natural objective for research, but is difficult to gauge the level of performance that is currently attainable. We describe a study that offers a rough benchmark. Speech data came from five passages of about 100 syllables each. They had been selected following pilot studies because they were effective at evoking specific emotion - fear, anger, happiness, sadness, and neutrality. 40 subjects were recorded reading them.

A battery of 32 potentially relevant features was extracted using our ASSESS system. They were broadly speaking prosodic, derived from contours tracing the movement of intensity and pitch. They were input to statistical decision mechanisms, of two types. Discriminant analysis uses linear combinations of variables to separate samples that belong to different categories. There are reasons to suspect that linear combination will not be appropriate, so neural net classifiers were also considered. An automatic relevance determination procedure was used to identify the most relevant parameters.

Discriminant analysis outperformed the neural networks. Using 90% of the data for training, and testing on the remaining 10%, a classification rate of 55 % (+/- 0.08%) was achieved. The most useful predictors covered a variety of properties – intensity (relative to the start of the passage) and its spread; pitch spread; durations of silences, rises in intensity, and syllables; and a property related to the shape of 'tunes', the number of inflections in the F0 contour per tune. Many more variables were less important, but nevertheless contributed.

1. INTRODUCTION

One of the natural goals for research on the vocal expression of emotion is automatic identification of a speaker's emotional state. The task is only beginning to be addressed, and the scarcity of published results makes it difficult to evaluate progress. This paper describes the level of success achieved in collaborations involving the team at Queen's University, Belfast. We regard it as a benchmark, which we expect our group and others will be able to surpass in the near future. The general outline of our work can be summarised as follows.

The speech that we have used involves passages about strongly emotive subjects written in words chosen to facilitate vocal expression of the emotions. We will call them emotive texts. Passages with varying verbal content are not appropriate for some kinds of research on emotion, but here they are a useful tool, partly because preprocessing eliminates most information that is related to word identity. They were so successful that some subjects could not bear to read the text that evoked sadness.

Preprocessing is carried out by a system called ASSESS (**A**utomatic **S**tatistical **S**ummary of **E**lementary **S**peech **S**tructures), which automatically recovers a range of measures from the speech signal. Most of the measures considered here are statistical summaries of properties related to prosody. Later versions of ASSESS measure a wider range of properties.

Statistical learning rules are then applied to ASSESS measures, giving rules that classifying speech samples into emotional categories.

Incorporated in that broad approach is a range of techniques that may be useful to groups working in the area.

Emotive texts are an interesting intermediate between truly spontaneous emotion and neutral sentences overlaid with emotional expression. Above all, they provide an opportunity to register evidence that lies on a relatively coarse temporal scale – for instance a crescendo, or a sustained rhythmic pattern. Research based on single sentences is effectively blind to effects at that level, and we are unconvinced that emotion can really be imposed on a neutral passage of any length. The material that we have analysed addresses those issues in a limited way, but we will report developments from it at the end of the paper.

ASSESS embodies one of the natural strategies in the area, that is, to develop systems that deliver automatically the widest possible range of measures that are likely to be relevant to the expression of emotion – paving the way for empirical evidence to dictate which measures are eventually used. The natural alternative is to use as few measures as possible, motivating the choice on theoretical grounds – a top-down strategy, in contrast to the bottom-up strategy that ASSESS represents. The best macro-strategy is probably for both approaches to be pursued in parallel, with the freest possible interaction.

Statistical learning is clearly the most promising approach to integrating evidence. What is not clear is whether familiar techniques are adequate, or whether theoretical innovations are needed. We report explorations of some less familiar techniques, and procedures that may be useful models for the study of alternative options.

2. THE SPEECH SAMPLE

2.1. Readers

The readers were 40 volunteers (20 males, 20 females) who were recruited from a GP surgery. They were required to: (1) be aged 18 to 69 years; (2) have no known speech or hearing impairment; (3) have no mental health problems (or psychiatric history) or learning disabilities; (4) have no head injury or brain trauma; and (5) have no respiratory tract problems or symptoms of the common cold.

2.2. Procedure

Eight texts of comparable length (7-8 sentences) were devised to convey anger, happiness, fear and sadness. They described highly emotive subjects in informal language appropriate to the emotion. Two emotionally neutral texts were also designed to provide a baseline. A pilot study with 20 volunteers identified five of the texts (one for each emotional state) which were always recognised as conveying the intended emotional state. In a subsequent study, 20 different raters considered the wording of the selected texts, and confirmed that (in their judgement) it was suited to expressing the relevant emotion.

Readers were asked to read each text aloud using the emotional expression which they felt was appropriate, in random orders. High quality recordings were obtained in quiet and echo-free surroundings using a professional Marantz recorder and dynamic AKG stationary microphone. The exercise produced 197 passages (three people declined to read the 'sad' passage which depicted the death of a father).

3 SPEECH VARIABLES

3.1. The ASSESS system

Speech samples were analysed using the ASSESS system. It generates a simplified core representation of the speech signal based mainly on the F0 and intensity contours. Key 'landmarks' are then identified, including peaks and troughs in the contours as well as boundaries of pauses and fricative bursts. Measuring the 'pieces' between these landmarks gives rise to a range of variables that we call 'piecewise'. They provide a rich description of the way contours (of pitch and intensity) behave over time. Variables, piecewise and others, are then summarised in an array of statistics (covering central tendency, spread and key centiles). Additional measures deal with properties of 'tunes' (i.e. segments of the pitch contour bounded at either end by a pause of 180 ms or more). For a fuller description, see Cowie, Sawey & Douglas-Cowie [1].

ASSESS is designed to recover some spectral properties, but the version used in this study was affected by an error in that area (involving variable typing). It introduced noise rather than systematic bias, but with one exception, it was judged best to set the affected measures of spectral properties aside. The problem has been corrected in later versions.

3.1. Preliminary selection of variables

Speech Measures which were obviously were content-related (e.g. related to passage length) were excluded at the outset, leaving 375 ASSESS measures to be considered. Finding effective ways to reduce that number is one of the keys to using ASSESS-type systems.

Measures relating to tunes	
1	tune duration
2	fit of tune to a quadratic function
3	no of inflections in F0 contour per tune
Spectral	
4	Energy below 250 Hz
Intensity contour (excluding pauses)	
5	Mean intensity
6	Median intensity
7	Inter-quartile range of intensity distribution
Intensity at local extrema in the intensity contour	
8	Mean at maxima
9	Inter-quartile range for intensities at maxima
10	Mean at minima
11	Inter-quartile range for intensities at minima
Magnitude of rises or falls in the intensity contour	
12	Inter-quartile range for magnitudes of rises
13	Inter-quartile range for magnitudes of falls
Pitch of points in the F0 contour	
14	Number of contributing observations
15	Mean
16	Inter-quartile range
Pitch at local extrema in the F0 contour	
17	Inter-quartile range for pitch at maxima
18	Inter-quartile range for pitch at minima
19	Inter-quartile range for pitch at all local extrema
Magnitude of rises in the F0 contour	
20	Median
21	Inter-quartile range
Durations of rises and falls in the intensity contour	
22	Median duration for rises
23	Median duration for falls
Durations of level sections in the intensity contour ('plateaux')	
24	Inter-quartile range for plateaux at intensity peaks
25	Upper limit (90%) of range for plateaux at intensity peaks
26	Median for plateaux at intensity minima
27	Inter-quartile range for plateaux at intensity minima
Durations of features in the F0 contour	
28	Median of silence durations
29	Inter-quartile range for durations of silences
30	Median duration of falls
31	Median duration of plateaux at F0 maxima
32	Inter-quartile range for duration of plateaux at F0 maxima

Table 1: ASSESS features used for classification

As a first step, one-way analysis of variance (ANOVA) followed by post-hoc comparisons of means (using Duncan's range test) was used to identify differences between each of the four passages and the neutral baseline. Measures were considered distinctive only if the overall ANOVA was significant at the $p<0.05$ level and the emotional passage contrasted with neutral passage with $p<0.05$.

More than a third of the ASSESS measures (n=136) produced differences between passages which were statistically significant. Sixty were spectral measures and, for the reasons stated earlier, all but one were excluded from the analysis. To reduce the remaining 75 measures, variables were selected as markers of emotion when: (a) they usually occurred on *more than one* passage (a measure occurring in only *one* passage is more likely to have occurred by chance alone); and (b) they were generally among the simplest measures shown (i.e. simple quantities such as overall level were preferred to more complex ones such as change in level, and simple measures of those dimensions, such as means and medians, were prefered to higher order measures, such as standard deviations and inter-quartile ranges).

32 variables in all were selected as robust markers of emotion using the procedures described above. Table 1 gives brief descriptions of them.

4: CLASSIFICATION

Analysis was restricted to the 32 selected features. First we compare the classification performances of different algorithms. Second, the best performing method was used to analyse the relevance of each input feature for the classification of emotions.

4.1. Methods considered

Three classification algorithms were contrasted and tested.

The first method is the default classifier, namely linear discriminants. In linear discriminants the classes are separated simply by linear planes. Logic suggests that non-linear relationships are likely to be important in principle (e.g. the relevance of one feature is conditional on others, or the ratio of features is indicative). However, more complex methods may be counterproductive if the data set is too small or too noisy.

The second method, Support Vector Machines (see Schölkopf et al. [2]), has empirically been shown to give good generalization performance on a wide variety of problems. In particular, SVMs show a competitive performance on problems where the data are sparse (many features, few data) and noisy as is the case with the ASSESS data-base.

In Support Vector Machines one has the freedom to choose a similarity measure, which is a function that determines how similar two data examples are. In this study we tested two of these measures, a linear and a Gaussian similarity measure.

The third method, Generative Vector Quantization, has been developed recently by the KUN group (Westerdijk et al. [3] and [4]). It has been shown that this method gives a comparable performance with state-of-the-art classifiers (sigmoid belief networks, wake-sleep algorithm) on handwritten digit recognition. It outperforms standard methods such as nearest neighbour and back-prop on this problem. The purpose of GVQ is to give a clear understandable representation of the structures that are present in the data. GVQ explains data examples by simple compositions of elementary features.

4.2. Performance of classification methods

The data set was randomly split up into 10 parts with which 10 experiments were performed. In each experiment one part out of the 10 was used as a test set while the other 9 parts were used as training data.

One of the attractions of nonlinear techniques is that information such as the speaker's gender could in principle be taken into account. To investigate that possibility, we performed experiments with and without the gender feature.

The resulting average classification scores are summarised in table 1. The uncertainty values in the table are the standard deviations over the 10 experiments. Three classification algorithms were contrasted and tested. Overfitting presumably explains why GVQ performs worse with 2 features than with 1.

	Test score (no gender)	Test score (gender)
Linear SVM	0.21 ± 0.05	
Gaussian SVM	0.52 ± 0.1	0.51 ± 0.1
GVQ, 1 feature	0.43± 0.1	0.43± 0.1
GVQ, 2 features	0.34 ± 0.09	0.36 ± 0.09
Linear Discriminants	0.55 ± 0.08	

Table 2: Classification methods and their performance.

Relevance of ASSESS features

The classification experiments show that the Linear Discriminant method gives the best performance. Therefore, we chose that method for the `feature relevance' experiments.

The data set was randomly split into 5 equal parts. Call this partition A. Each of the 5 parts was successively used as an independent test set. Each 4/5 subset of partition A was again split up into 5 subparts. This is partition B. On the 4/5 subset of partition B we trained 32 models for each of which we omitted one of the features. The scores were then compared on the remaining 1/5 part of partition B. The least relevant feature (this is the feature for which the score dropped the least) was then deleted from the feature set. To test the generalization performance, the corresponding best performing classifier was in addition tested on the 1/5 subset of partition A. With the 31 remaining features we trained 30 models on the same 4/5 part of partition B again omitting one feature for each classifier. The models were compared again and the least important feature was deleted from the feature set. We continued in this manner until only one feature remained (the most relevant feature).

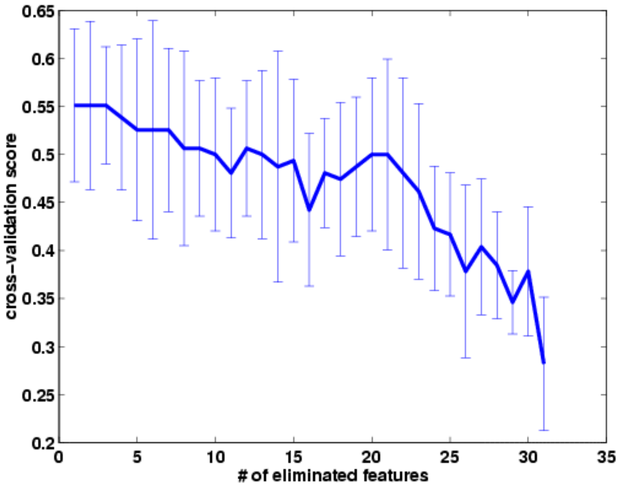

Figure 1: Test set scores as a function of the number of eliminated input features.

14	Number of F0 points recovered
11	Inter-quartile range for intensities at minima
28	Median of silence durations
4	Energy below 250 Hz
10	Mean intensity at intensity minima
3	No of inflections in F0 contour per tune
22	Median duration for rises in intensity
16	Inter-quartile range: Pitch of points in the F0 contour
13	Inter-quartile range for magnitudes of falls in intensity
18	Inter-quartile range for pitch at F0 minima
12	Inter-quartile range for magnitudes of intensity rises
23	Median duration for intensity falls
31	Median duration of plateaux at F0 maxima
9	Inter-quartile range for intensities at intensity maxima
30	Median duration of falls in F0
15	Mean pitch of points in the F0 contour

Table 3: Features making the highest contributions to classification by discriminant analysis

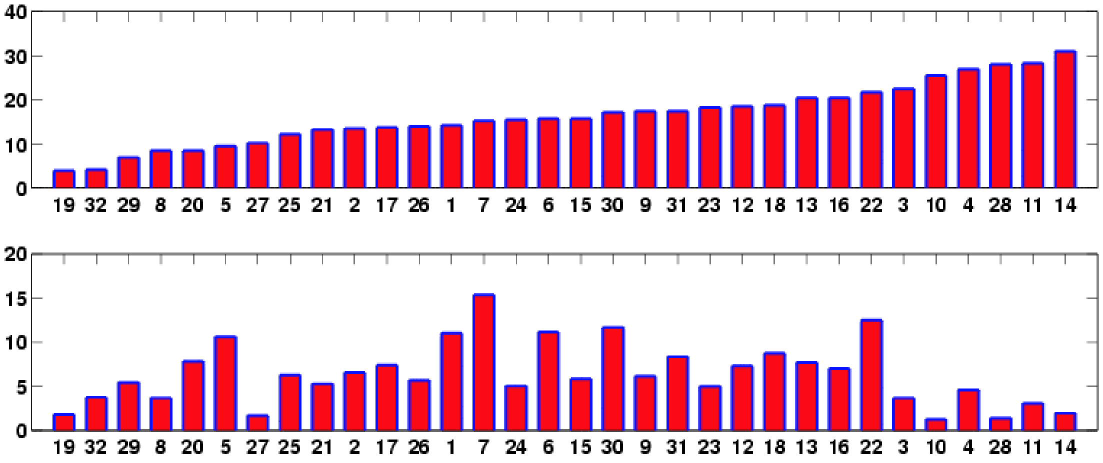

Figure 2: The average ranking of each feature. The feature that was deleted last has the highest ranking (32). The upper panel shows mean ranking, and the lower shows the standard deviation of ranking.

The whole experiment was repeated 5 times, training models on each 4/5 subset of partition A. Hence, in total we trained 5*(32 + 31+ 30 + … + 1)=5*32*32/2=2560 classifiers which took about 24 hours of CPU time. The test set scores (the scores on the 1/5 parts of partition A) at each stage of feature exclusion are shown in figure 1

Figure 2 shows the average ranking of each feature. The feature that was deleted last has the highest ranking (32). Since the experiment was performed five times we could compute the standard deviations of the rankings. These are plotted in the lower part of figure 2. Table 3 identifies the 16 features that contributed most to classification, preceded for purposes of cross-reference by the identifying numbers used in table 1.

It was recognised after the main analysis that one of the variables, number of F0 points recovered, reflected the fact that the texts contained different numbers of syllables. Because its contribution is large, an additional discriminant analysis was run with the original variable replaced by one that had been normalised with respect to syllable length, using a 1 in 10 split as in Table 2. Recognition on the test set averaged 52.3%, indicating that performance was lowered, but remained of the same general order. Removing the variable altogether reduced recognition rate again to 49%.

Confusion patterns were generally straightforward. Table 4 shows the confusions on the test sets (using the normalised number of F0 points measure). They were relatively evenly distributed across categories.

Intended emotion	Classification generated				
	afraid	happy	neut	sad	angry
Afraid	20	4	4	4	8
Happy	4	21	9	4	1
Neutral	5	7	18	6	4
Sad	5	4	4	23	0
Angry	9	4	6	1	20

Table 4: Confusions on the test sets

Figure 3 shows how means varied from emotion to emotion in the eight highest ranking variables. A useful pointer to the way they distinguish the emotions is to consider pairs of emotions, and to note that for most pairs, there is at least one measure on which they are sharply distinct.

It is noticeable that although the variables in Figure 3 are all broadly speaking prosodic, ASSESS piecewise measurement allows a considerable range of relevant properties to be identified and recovered automatically.

Two measures dealt with intensity level, median intensity at intensity minima and energy below 250Hz. The effects that they reveal (in fear, happiness, and anger) take the form of a 'crescendo', because ASSESS uses the beginning of a passage to set an intensity scale for what follows. It is interesting that the strongest statistical effects come from selective measures of intensity, one measuring intensity in the lower spectrum and the other measuring intensity at instants where it reaches a local minimum. The effects ought to be confirmed before speculating on possible explanations. It is suggestive, though, that the most relevant measure of intensity spread also involved the variation associated with instants where intensity reached a local minimum. Note that anger was associated with high intensity, but low variation.

Two measures related to the F0 contour. One, unsurprisingly, was pitch spread. Less familiar was a property related to the shape of 'tunes', the number of inflections in the F0 contour per tune. It is a feature of piecewise measurement that variables of that kind can be recovered automatically.

Another feature of piecewise measurement is that a range of timing-related variables can be recovered. Three figured among the strongest predictors - durations of silences, rises in intensity, and number of F0 points recovered per syllable. The last is effectively a measure of syllable duration. Recovery is not truly automatic, because syllables were counted by hand. However, given that the measure seems to be important, approaches to automating it can be imagined.

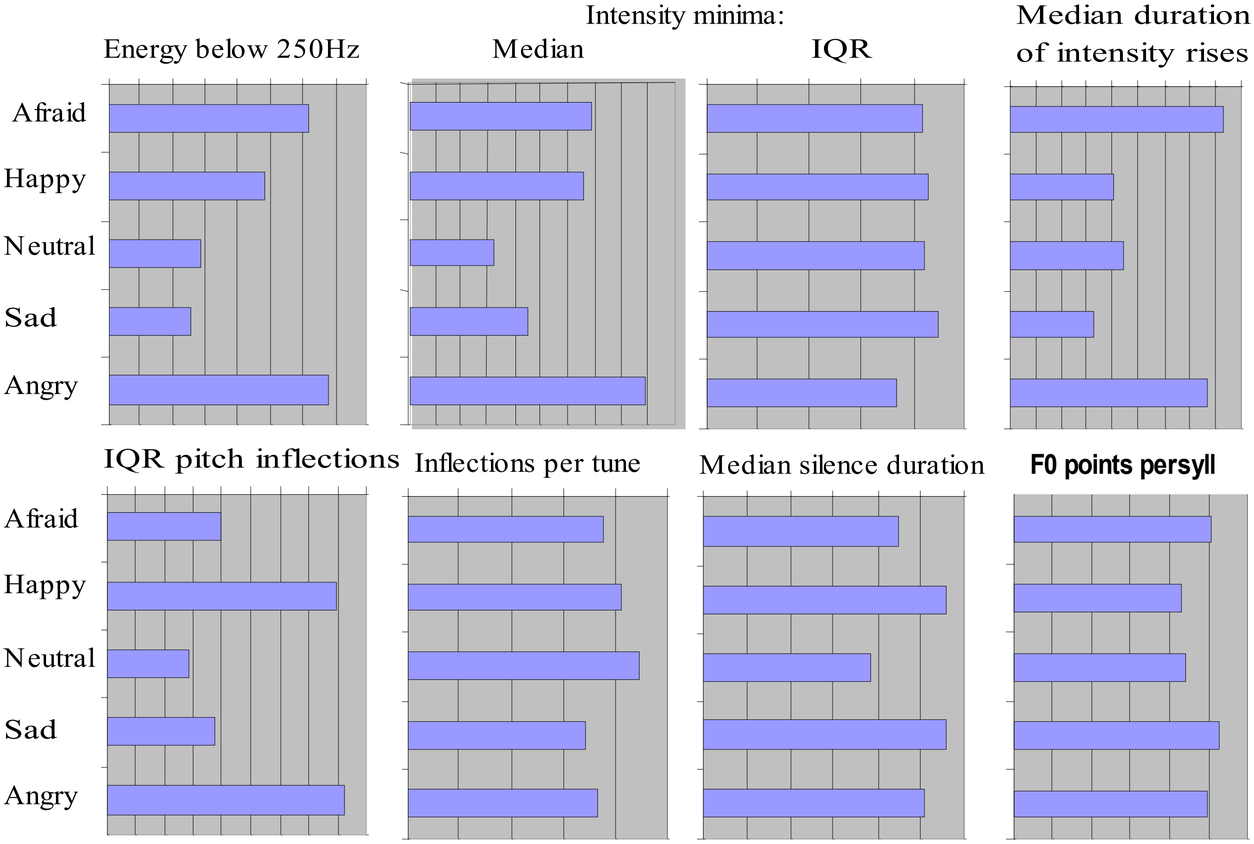

Figure 3: Profiles of different emotional states with respect to the eight ASSESS variables that contributed most to prediction

5. DISCUSSION

As a rough benchmark, our work suggests that 50% correct classification is an attainable goal for automatic systems aiming to discriminate among five emotional states. Informal reports from other groups suggest that attainable classification rates rise as the number of states to be discriminated falls

Banse and Scherer [5] have reported a similar level of classification to ours, but with a larger number of emotions. The comparison suggests that classification rate also reflects what might be called the purity of the sample. Their sample was constructed to guarantee sharp separation among types of speech; the expressions of emotion that they recorded seem likely to have been intense; and the utterances were short, so that the differences were concentrated. In contrast, our passages contained gradations of emotion; and they were longer, so that information at expressive peaks was diluted.

The use of less pure material relates to application. Our study originated in an applied project, concerned with distinguishing between people with normal ability to express emotion, and clinical groups where that ability is impaired [6]. Hence it had to work with signs of emotion that emotionally normal people in a clinical setting could be expected to generate, however impure they might be. What we describe is the kind of success rate that might be expected in that application. It is not clear how rates based on purer samples transfer to applied settings.

On a more technical level, our work makes a number of points. The differences between learning rules were large and unexpected. It is clear that the area is one where serious effort needs to be invested.

We have described techniques that seem appropriate to gauging how much features contribute to classification. An unexpected outcome was the sheer number of contributing features. It is natural for research in a linguistic tradition to assume that the task is to identify a small set of features that are highly relevant. Our findings suggest that the alternative ought to be taken seriously. Optimising discrimination may rest on using the widest possible range of relevant features.

A major issue for learning-based approaches is the scale of the database. Our sample contained 200 passages. It is clear that given the noise and complexity in the data, far larger samples are needed if advantage is to be taken of any but the most basic statistical learning techniques. The point applies equally to research that uses statistical tests. Very large samples are needed to draw statistical inferences of any subtlety. For the smaller samples that are usually considered, credible inference cannot be purely statistical: prior knowledge has to be invoked.

ASSESS emerges as a reasonable prototype for feature extraction in research on speech varieties. The unintended absence of spectral information had the benefit of underlining the variety and usefulness of the prosody-related variables that can be extracted automatically from a speech sample.

It is reassuring that though the approach avoids preconceptions about relevant features, so many high-ranking features were familiar in a broad sense – intensity and its spread, pitch spread, and durations of silences and syllables. However, it is noticeable that the most useful measures of these attributes were not necessarily the most obvious. The main example is the way specialised measures of intensity prove more useful than basic ones. Also noticeable is the fact that some much less standard measures contributed strongly – notably the duration of rises in intensity, and 'tune shape' measured by the number of inflections per tune. It appears that open ended exploration of potentially relevant features is a strategy worth pursuing.

A number of developments pointed by these findings are currently under way. An improved version of ASSESS has been developed. It corrects the problem with spectral measurement. More radically, it extends the use of tune-type units. Having identified units bounded by appreciable silences at either end, the system now presents statistical summaries for each tune. One aim is to increase the number of of inputs available to statistical learning algorithms A typical passage contains a dozen tunes or so, and averaging over them is probably throwing away information. The second aim is to give algorithms a chance to identify expressive peaks, rather than masking them in more general background material.

The second aim depends on obtaining information about variation of emotional signs across a passage rather than assigning the whole to a single emotion category. We have developed the Feeltrace system, which we describe elsewhere in this conference, as a way of doing that.

We have also extended the strategy of using emotive texts, selecting texts where raters judge that the phrasing lends itself to expressing the emotion vocally. That strategy offers the possibility of pinpointing higher order prosodic correlates that are precluded by brief samples and obscured by multiple other variables in truly natural databases.

6. REFERENCES

1. Cowie, R., Sawey, M. and Douglas-Cowie, E. (1995) A new speech .analysis system: ASSESS *Proc Internat Conf Phonetic Sciences*, 3, 278-281, Stockholm, 1995

2. B. Schölkopf, C.J.C. Burges, A. J. Smola. (1999) *Advances in Kernel Methods: Support Vector Learning,* MIT Press, Cambridge MA.

3. M.Westerdijk, D. Barber, W. Wiegerinck. (1999) Generative Vector Quantisation, *ICANN 99*, 2 934-939.

4. M. Westerdijk, and W. Wiegerinck, (2000) Classification with Multiple Latent Variable Models using Maximum Entropy Discrimination*, Proc. 17th Internat Conf on Machine Learning*

5. Banse, R. & Scherer, K. (1996) Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* 70 (3), 614-636.

6. Cowie, R. & Douglas-Cowie, E. (1996) Automatic statistical analysis of the signal and prosodic signs of emotion in speech. *Proc ICSLP 1996,* 1989-1992.