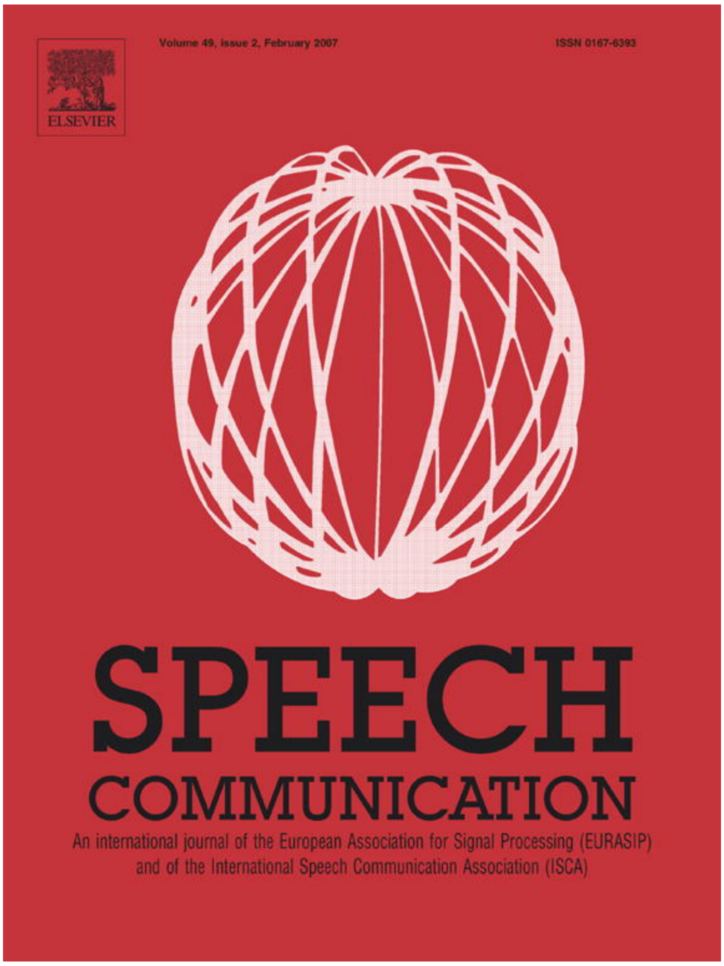
Volume 49, issue 2, February 2007
ISSN 0167-6393
ELSEVIER
SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP)
and of the International Speech Communication Association (ISCA)

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 98–112
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Ensemble methods for spoken emotion recognition in call-centres

Donn Morrison, Ruili Wang *, Liyanage C. De Silva

Institute of Information Sciences and Technology, Massey University (Turitea), Palmerston North, Private Bag 11222, New Zealand

Received 26 April 2006; received in revised form 3 October 2006; accepted 15 November 2006

Abstract

Machine-based emotional intelligence is a requirement for more natural interaction between humans and computer interfaces and a basic level of accurate emotion perception is needed for computer systems to respond adequately to human emotion. Humans convey emotional information both intentionally and unintentionally via speech patterns. These vocal patterns are perceived and understood by listeners during conversation. This research aims to improve the automatic perception of vocal emotion in two ways. First, we compare two emotional speech data sources: natural, spontaneous emotional speech and acted or portrayed emotional speech. This comparison demonstrates the advantages and disadvantages of both acquisition methods and how these methods affect the end application of vocal emotion recognition. Second, we look at two classification methods which have not been applied in this field: stacked generalisation and unweighted vote. We show how these techniques can yield an improvement over traditional classification methods.

Keywords: Affect recognition; Emotion recognition; Ensemble methods; Speech processing; Speech databases

1. Introduction

With the ubiquity of automated systems in today's society comes the burden of interacting with these systems due to the lack of machine-based emotional intelligence. For example, the emotional information conveyed through speech is an important factor in human–human interaction and communication. Humans feel most natural communicating with other humans because the extra information represented in their emotional expressions can be recognised, processed, and reflected. Hence, when humans interact with computer systems, there is a gap between the information conveyed and the information perceived.

Emotional intelligence is defined by Salovey et al. (2004) as having four branches: the perception of emotion, emotions facilitating thought, understanding emotions, and managing emotions. The work in this study is dedicated to the *perception* of human emotion from the prosodic properties of speech. In other words, this study aims to build a system that can capture and interpret the vocal expression of emotion in humans. More specifically, we seek to improve on traditional emotional speech classification methods using ensemble or multi-classifier system (MCS) approaches. We also aim to examine the differences in perceiving emotion in human speech that is derived from different methods of acquisition.

In this study we also look at applications of emotionally intelligent systems in call-centres (see Fig. 1). Call-centres often have a difficult task of managing customer disputes. Ineffective resolution of these disputes can often lead to customer discontent, loss of business and in extreme cases, general customer unrest where a large amount of customers move to a competitor. It is therefore important for call-centres to take note of isolated disputes and effectively train service representatives to handle disputes in a way that keeps the customer satisfied (Petrushin, 2000).

Automated telephone systems are another potential application area that humans find themselves interacting with more and more. These systems have speech recognition units that process user requests through spoken

[*] Corresponding author.
E-mail addresses: d.morrison@massey.ac.nz (D. Morrison), r.wang@massey.ac.nz (R. Wang).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.11.004

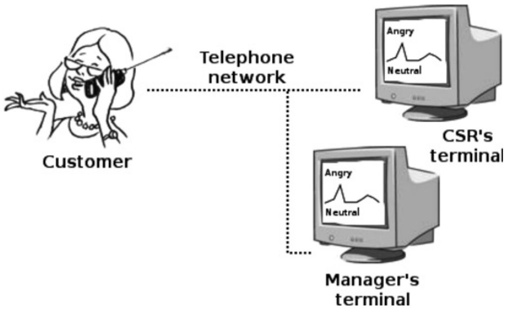

Fig. 1. Affect recognition in a call-centre environment.

language. A spoken affect recognition system can help process calls according to perceived urgency. If a caller is detected as being angry or confused in the automated system, their call can be switched over to a human operator for assistance. This could be particularly useful for the elderly who can often be disoriented when interacting with automated telephone systems. In (Petrushin, 2000) a system was built to monitor voice-mail messages in a call-centre and prioritise them with respect to emotional content. Similarly, in (Liscombe et al., 2005), prosodic and contextual features were used to identify emotion in a spoken dialog system. Such systems can make interaction with automated call-centres more efficient and less daunting.

Many machine learning algorithms have been applied to the problem of automatic emotion detection from speech. However, very few studies consider hybrid or ensemble classification methods. In (Petrushin, 2000), ensembles of neural networks were employed to improve classification accuracy on a two-class problem. A rate of 75% was achieved. In (Dellaert et al., 1996), classifiers were grouped together based on cooperation with others in a set. A majority vote yielded the final prediction, and for a five-class problem a classification accuracy rate of 79.5% was seen. In (Nakatsu et al., 1999), an ensemble of eight neural networks, one trained for each emotion, was devised. An accuracy rate of 50–55% was achieved.

We propose two existing ensemble classification methods which are new to the field. The first, stacked generalisation, employs several base-level classifiers to get class predictions which are then used by a meta-learning algorithm during the training phase in an attempt to predict when the base-level classifiers are incorrect (Wolpert, 1992). The second, a variation on majority voting, is a collection of unique classifiers, each trained on the same data, with predictions combined in an unweighted voting scheme. Both approaches are discussed in detail in Section 6.

The rest of the paper is organised as follows. Section 2 first introduces methods of data acquisition and then presents the data sets used in the study. Relevant speech features, which have been shown to correlate to emotional states, are reviewed in Section 3. Next, Section 4 introduces a set of prosodic features which forms a basis for the extraction of discriminative emotional information. Several traditional classification methods and feature selection techniques are then described in Section 5. Section 6 introduces the two improvements over the traditional classification methods: an unweighted voting scheme and stacked generalisation. Section 7 shows the results of classification experiments and provides a discussion. Finally, in Section 8 we conclude and offer directions for future research.

2. Data acquisition

There are three methods of data acquisition in emotion research (Scherer, 2003). The first is natural expression, where data is collected from a real-world situation where users are not under obvious observation and are free to express emotions naturally, as they would in an everyday situation. The second is induced emotional expression, where naïve users are presented with scenarios that induce the required emotional response. Last, speech acquisition using simulated or portrayed emotional expression makes use of professional or non-professional actors and actresses. Subjects are instructed to produce emotional expressions for various emotion classes, with varying degrees of intensity or arousal.

The two data sets used in this study are summarised in Table 1. The first data set is taken from a natural scenario and was initially the primary focus of the research. Later, a second data set was acquired from non-professional actors and actresses portraying emotional speech for the purpose of validating the features and algorithms used on the first data set. The second data set was subsequently integrated into the study in order to compare useful features and properties against the first data set.

2.1. Natural data collected from a call-centre

The first database used in this research was provided by a call-centre that handled customer inquiries for several

Table 1
Summary of the data sets used in this study

Name	Description	Emotion classes	Speakers	No. of utterances
NATURAL	Collected from a call-centre for an electricity company	2 (anger, neutral)	11	388
ESMBS	Collected from Burmese and Mandarin non-professional actors	6 (anger, happiness, sadness, disgust, fear, surprise)	12	720

The NATURAL data set is collected from a call-centre and the ESMBS data set is obtained from a previous study and consists of utterances by non-professional actors and actresses.

electricity companies. Customers call and speak directly to a customer service representative (CSR). The customers query or provide information about their accounts, billing information, address, payment methods, etc. Often, customers have a dispute to resolve with the company and subsequently, emotions are expressed.

For this data set, the average length of a conversation between a customer and CSR was 3 min and 40 s. The median call length was 2 min and 38 s. The longest call duration was 34 min and 3 s, and the shortest recordings were around 1 s, but involved no audible speech and were probably the result of some technical error during the recording process.

The data consisted mainly of neutral speech, with the second largest category being angry speech. Table 2 shows the distributions for respective emotion classes. Because of the low distributions of happiness, sadness, fear, disgust, and surprise, it can be assumed that the probability of these occurring in the call-centre are quite low, and because of this it is safe to consider only anger and neutral emotional states. The small amounts of data collected under the other emotional states (happiness, sadness, surprise, fear, disgust), were excluded. Similarly, (Devillers et al., 2002) also used data from a customer service centre. This study also found low emotion distribution and subsequently retained two of the basic emotion classes, anger and fear, because the probabilities of other emotions in that context were very low. In (Ang et al., 2002), used induction methods for collecting emotional speech data and observed a high amount (84%) of neutral samples, followed by a low amount (8%) of annoyance. Due to this they limited their study to include only annoyance and frustration versus everything else.

Each conversation file was manually segmented into phrase-level utterances selected from a total of 11 speakers, 2 male and 9 female. The maximum utterance duration is 13.60 s, the minimum is 0.52 s, and the mean duration is 3.25 s. Each utterance was stored in 16 bit PCM WAVE format sampled at 22,050 Hz.

Initially, the data set comprised 190 angry utterances and 201 neutral utterances, totalling 391. To gain an objective ground truth, nine listener-judges were instructed to classify the entire data set according to the predefined class labels. The listener-judges yielded a mean agreement of 81.95% and the final data set comprised 155 angry utterances and 233 neutral utterances. In total there were 388 utterances (three utterances were labelled as ties and were subsequently discarded). This data set is labelled NATURAL.

2.2. Simulated data from the ESMBS database

The ESMBS database (Emotional Speech of Mandarin and Burmese Speakers) was collected for a previous study on emotion recognition (Nwe et al., 2003). This data set was collected to study emotional effects on vocal parameters.

The data set was collected from a set of 12 non-professional actors and actresses. Six Mandarin and six Burmese speakers were used, each speaking their native language, with each of these six consisting of three men and three women. Each speaker recorded ten different utterances for each of the six emotions. In total, for the 12 speakers, there were 720 emotional utterances.

The emotion set represented by this data set are the six prototypical emotions most often studied in this field: anger, disgust, fear, joy, sadness, and surprise. The mean length of the samples in the data set was 1.50 s. All speech samples were recorded with 16 bits per sample at 22,050 Hz and stored in PCM WAVE format. The content of each utterance was a phrase or sentence which also contained emotion from one of the above six.

Four listeners were used to judge the emotional content of each utterance. These listeners could not understand the language of the respective speakers, so vocal characteristics and not contextual information was the only information used for classification. Average classification accuracy by human evaluation was found to be 65.7% (68.3% for Burmese and 63.1% for Mandarin). These figures coincide with previous studies (Dellaert et al., 1996; Petrushin, 2000; Polzin and Waibel, 2000), as well as previous research on cross-cultural emotion recognition from speech (Scherer, 2003; Elfenbein and Ambady, 2002) which typically describe human classification rates between 55% and 70%.

3. Properties of emotional speech

Prosodic parameters have been found to represent the majority of emotional content in verbal communication (Murray and Arnott, 1993; Scherer, 2003). Of these, fundamental frequency (pitch), energy, and speaking rate are widely observed to be the most significant characteristics (Batliner et al., 2003; Lee et al., 2004; Dellaert et al., 1996; Ang et al., 2002; Huber et al., 2000; McGilloway et al., 2000; Polzin and Waibel, 2000; Nwe et al., 2003). The specific correlations between the basic emotions and these prosodic features are discussed below.

Table 2
Distribution of perceived speaker affect from natural corpus (NATURAL)

Number of conversations (%)	Emotion class
93.3	Neutral
3.1	Anger
1.8	Happiness
0.1	Sadness
0.0	Surprise, fear, disgust

3.1. Fundamental frequency and emotional speech

The fundamental frequency (F0), often referred to as the pitch, is one of the most important features for determining emotion in speech (Nakatsu et al., 1999; Polzin and Waibel, 2000; Petrushin, 2000; Ang et al., 2002; Lee et al., 2004). The fundamental frequency is defined as the lowest frequency at which the speech signal repeats itself (O'Shaughnessy, 2000).

The F0 contour has been shown to vary depending on the emotional state being expressed. Cowan (xxx) discovered that neutral or unemotional speech has a much narrower pitch range than that of emotional speech, and found that as the emotional intensity is increased, the frequency and duration of pauses and stops normally found during neutral speech are decreased (Murray and Arnott, 1993).

More specifically, angry speech typically has a high median, wide range, wide mean inflection range, and a high rate of change (Fairbanks and Pronovost, 1939). Williams and Stevens (1972) discovered vowels of angry speech to have the highest F0, and (Fonagy, 1978) found that angry speech exhibits a sudden rise of F0 in stressed syllables and the F0 contour has an "angular" curve. Frick (1986) postulated that frustration, which has similar but less extreme physiological causes as anger, has a higher fundamental frequency than neutral speech. Scherer (1996) describes anger as having "an increase in mean pitch and mean intensity." Downward slopes are also noted on the pitch contour. Breazeal and Aryananda (2002) found that prohibitions or warnings directed at infants are spoken with low pitch and high intensity in "staccato pitch contours." Cowan (xxx) and Fonagy and Magdics (1963) found that happiness expressed in speech, like anger, has an increased pitch mean and pitch range.

Fear was discovered to have a high pitch median, wide range, medium inflection range, and a moderate rate of change (variation) (Fairbanks and Pronovost, 1939; Williams and Stevens, 1972), and increased pitch level is also apparent (Fonagy, 1978). Conversely to fear exhibiting a wide range, there are reports that fear instead has a narrow F0 range (Fonagy and Magdics, 1963).

Contrasting these more excited emotions are sadness and disgust which typically have lower physiological activation levels. Sadness is shown to yield lower pitch mean and narrow range (Skinner, 1935; Davitz, 1964; Fonagy, 1981; Oster and Risberg, 1986; Johnson et al., 1986). Fairbanks and Pronovost (1939) report that disgust generally has a low pitch median, wide range, lower inflectional range, lower rate of pitch change during inflection. As with fear, there are contrasting findings with (Fonagy and Magdics, 1963) noting disgust having a narrow pitch range.

Fig. 2 shows the pitch contours of two example utterances from the NATURAL data set. It can be seen that the angry sample has downward slopes, concurring with (Scherer, 1996), and a greater range. The neutral sample has a monotonous contour with a shallow range.

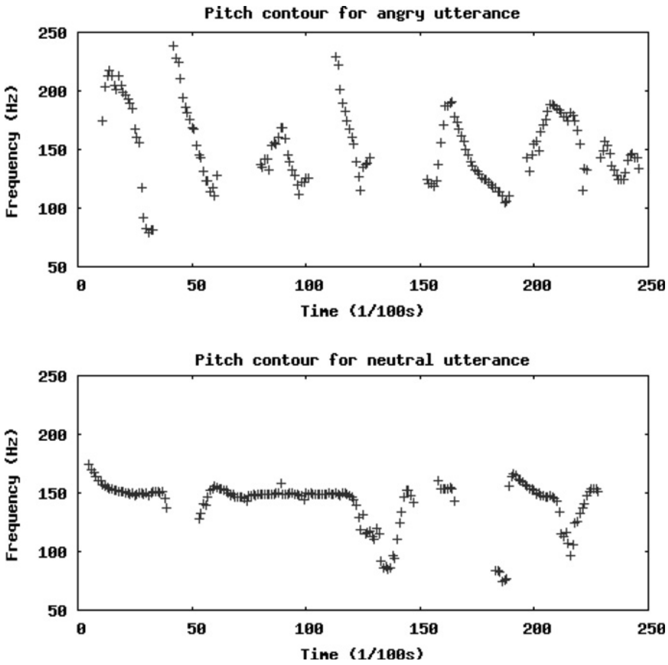

Fig. 2. Example pitch contours for anger and neutral utterances from the NATURAL data set. The contour for angry speech typically has a much wider range, while neutral speech is narrow and monotonous.

3.2. Formant frequencies and emotional speech

The resonant frequencies produced in the vocal tract are referred to as formant frequencies or formants (Rabiner and Schafer, 1978). Although some studies in automatic recognition have looked at the first two formant frequencies (F1 and F2) (Petrushin, 2000; Lee et al., 2004), the formants have not been extensively researched.

Williams and Stevens (1972) found that anger produced vowels "with a more open vocal tract" and from that inferred that the first formant frequency would have a greater mean than that of neutral speech. It was also noticed that the amplitudes of F2 and F3 were higher with respect to that of F1 for anger and fear compared with neutral speech. Neutral speech typically displays a "uniform formant structure and glottal vibration patterns," contrasting the "irregular" formant contours of fear, sadness, and anger.

Scherer (2003) lists predictions of the formant frequencies along with several emotion classes. For happiness, it is noted that the F1 mean is decreased while the F1 bandwidth is increased. For anger, fear, and sadness, the F1 mean is increased while the F1 bandwidth is decreased. F2 mean is decreased for sadness, anger, fear, disgust.

3.3. The use of energy as an emotional marker

Energy, often referred to as the volume or intensity of the speech, is also known to contain valuable information (Huber et al., 1998; Nakatsu et al., 1999; Polzin and Waibel, 2000; McGilloway et al., 2000). The intensity contour provides information that can be used to differentiate sets of emotions.

In their research, (Fonagy, 1981) found that angry speech had a noticeably increased energy envelope. Happiness showed similar characteristics, as reported by Davitz (1964); Skinner (1935). Sadness was associated with decreased intensity (Fonagy, 1981; Davitz, 1964) and disgust had reduced loudness (Fonagy and Magdics, 1963). Scherer (2003) notes that in fear, joy, and anger there is an increase in high frequency energy, whereas sadness has a decrease in high frequency energy.

These characteristics follow with what is expected of the emotional state. Those with high activation levels such as anger, surprise, and happiness generally have a higher intensity, while fear, sadness, and disgust have lower intensity (Nwe, 2003).

3.4. Rhythm-based characteristics

Properties of rhythm-based characteristics include pauses between voiced sounds, lengths of voiced segments, and rate of speech (articulation). The rate of speech is usually calculated by measuring the number of syllables per second.

Speaking rate has been used in previous research (Dellaert et al., 1996; Huber et al., 1998; Petrushin, 2000; Ang et al., 2002). It has been noted that fear, disgust, anger, and happiness often have a higher speaking rate, while surprise has a normal tempo and sadness a reduced articulation rate (Nwe, 2003).

Fairbanks and Hoaglin (1941) and Fonagy (1981) found that anger has an increased speech rate, and "pauses forming 32% of total speaking time." Happiness has been shown to have anywhere from a slower tempo (Oster and Risberg, 1986), to a "regular" rate (Davitz, 1964), to even an increased rate (Fonagy, 1981). For sadness, on the other hand, it has been generally agreed that the tempo is slower (Skinner, 1935; Davitz, 1964; Fonagy, 1981; Oster and Risberg, 1986; Johnson et al., 1986) and the speech contains "irregular pauses" (Davitz, 1964).

Williams and Stevens (1972) stated that fear exhibited a reduced speech rate, while (Fairbanks and Hoaglin, 1941) contrasts this by noting a high speech rate, and "pauses forming 31%." Disgust has a very low speech rate, increased pause length, with pauses typically comprising 33% of speaking time (Fairbanks and Hoaglin, 1941). The correlations mentioned above are summarised in Table 3.

4. Prosodic features

Based on the acoustic correlates described in the previous section and the literature relating to automatic emotion detection from speech, we selected features based on four prosodic groups: *the fundamental frequency, energy, rhythm*, and *the formant frequencies*. The fundamental frequency, energy, and formant frequencies are represented as contours. From these contours, we selected seven statistics: *the mean, minimum, maximum, standard deviation, value at the first voiced segment, value at the last voiced segment*, and *the range*. For the rhythm-based features, we selected three: *the speaking (articulation) rate, average length of unvoiced segments (pause)*, and *the average length of voiced segments*.

In total, we selected 38 prosodic features which are used as a starting point for describing the variation between angry and neutral speech. These are listed in Table 4.

For the extraction of the pitch contour, we used the Robust Algorithm for Pitch Tracking (RAPT) (Talkin, 1995). This algorithm uses the cross-correlation function to identify pitch candidates and then attempts to select the "best fit" at each frame by dynamic programming. One of the benefits of using the cross-correlation function is that it does not suffer the windowing dilemma of the autocorrelation function while maintaining resolution for high pitch values and the ability to detect low pitch values (Rabiner and Schafer, 1978).

The first three formant frequencies were extracted using linear predictive coding (LPC) and dynamic programming to select optimal candidates based on their scores in relation to previous candidates. The candidates are then ranked according to their relative location, bandwidth, and relation to the previous formant candidates. The best candidates are selected for each formant using dynamic

Table 3
Speech correlations of the basic emotions

	F0 mean	F0 range	Energy	Speaking rate	Formants
Anger	Increased	Wider	Increased	High	F1 mean increased; F2 mean higher or lower, F3 mean higher
Happiness	Increased	Wider	Increased	High	F1 mean decreased; F1 bandwidth increased
Sadness	Decreased	Narrower	Decreased	Low	F1 mean increased; F1 bandwidth decreased; F2 mean lower
Surprise	Normal or increased	Wider	–	Normal	–
Disgust	Decreased	Wider or narrower	Decreased or normal	Higher	F1 mean increased; F1 bandwidth decreased; F2 mean lower
Fear	Increased or decreased	Wider or narrower	Normal	High or low	F1 mean increased; F1 bandwidth decreased; F2 mean lower

Table 4
Feature groups and statistics used for measuring differences between angry or neutral speech

Feature group	Statistics
Fundamental frequency (F0)	(1) mean, (2) minimum, (3) maximum, (4) standard deviation, (5) value at first voiced segment, (6) value at last voiced segment, (7) range
Formant frequencies (F1, F2, F3)	(8, 15, 22) mean, (9, 16, 23) minimum, (10, 17, 24) maximum, (11, 18, 25) standard deviation, (12, 19, 26) value at first voiced segment, (13, 20, 27) value at last voiced segment, (14, 21, 28) range
Short-time energy	(29) mean, (30) minimum, (31) maximum, (32) standard deviation, (33) value at first voiced segment, (34) value at last voiced segment, (35) range
Rhythm	(36) speaking rate, (37) average length of unvoiced segments (pause), (38) average length of voiced segments

Features are numbered in parentheses.

programming similar to that used for the RAPT (Rabiner and Schafer, 1978).

The energy envelope consists of the magnitude of the signal calculated over a frame or window in order to average or smooth the contour. The energy frame size should be long enough to smooth the contour appropriately but short enough to retain the fast energy changes which are common in speech signals and it is suggested that a frame size of 10–20 ms would be adequate (Rabiner and Schafer, 1978). In this paper we used a frame size of 10 ms.

The rhythm-based statistics are all based on the voiced and unvoiced segment durations. The rate of speech (articulation) is measured as the number of syllables normalised by the utterance duration. A syllable can be roughly defined as the transition from a voiced to unvoiced segment (one or more consecutive frames), or vice versa. A segment is deemed to be voiced if it is periodic, in other words if it has a value greater than zero for the fundamental frequency. A segment is unvoiced if it is aperiodic, or has no fundamental frequency.

5. Classification techniques

Classification was performed using WEKA (Waikato Environment for Knowledge Analysis).[1] WEKA is a data mining workbench that allows comparison between many different machine learning algorithms. In addition, it also has functionality for feature selection, data pre-processing, and data visualisation.

The selection of base-level classifiers was done by evaluating several algorithms over the NATURAL data set and selecting the top performers. As noted in Section 1, the NATURAL data set was used to determine the top base-level classifiers and the ESMBS data set was used to validate the choice of features and classifiers, as well as to compare important features against the NATURAL data set. It is this distinction that affords the selection of certain parameters to be obtained using only the NATURAL data set.

Table 5
Initial ranking of base classification algorithms on the NATURAL data set

Algorithm	Accuracy (%)
SVM (RBF)	76.93
KNN ($K = 5$)	75.85
Multi-layer perceptron	74.25
Random forest	71.98
K^*	70.67
Naive Bayes	69.56
SVM (polynomial)	69.50
C4.5 decision tree	67.47
Random tree	60.05

Table 5 shows the classification accuracies for the algorithms initially selected. In order to retain some degree of simplicity, only the top five algorithms are retained. As can be seen, the top performers are the support vector machine (SVM) with the radial basis function (RBF) kernel, the random forest, the multi-layer perceptron (artificial neural network), K^*, and K-nearest neighbour with $K = 5$. For the SVM, the use of the RBF kernel showed a significant improvement over the use of the polynomial kernel.

5.1. Support vector machines

Support vector machines (SVMs) are a relatively new machine learning algorithm introduced by Vapnik (1995). They are based on the statistical learning theory of structural risk management (SRM) which aims to limit the empirical risk on the training data and on the capacity of the decision function. Support vector machines are built by mapping the training patterns into a higher dimensional feature space where the points can be separated using a hyperplane.

In WEKA, SVMs are implemented as the sequential minimal optimisation (SMO) algorithm (Platt, 1998). There are two kernels available: polynomial, and radial basis function (RBF). As shown in Table 5, RBF performed better on our data set. The RBF kernel is defined as

$$K(x_i, y_j) = \exp(-\gamma \|x_i - y_j\|^2), \quad \gamma > 0 \tag{1}$$

Optimal values for the width of the RBF function, γ, and the cost parameter C, can be found by performing a grid search on the training data. For our experiments, a grid search of the training data yielded optimal values $\gamma = 0.7$ and $C = 8.0$.

5.2. Random forests

Random forests, invented by Breiman (2001), consist of ensembles of tree predictors. These tree ensembles are a

[1] http://www.cs.waikato.ac.nz/~ml/weka/.

method of growing a "forest" of decision trees by selecting features for each node randomly and independently of every other tree but with the same distribution. When a random forest has been grown, classification requires that the predictions of each tree are combined by voting to determine the overall prediction.

In (Breiman, 2001), the authors state that if we let $h_1(x), h_2(x), \ldots, h_k(x)$ be an ensemble of classification trees with random training vector Y, X, then the margin is defined as

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \tag{2}$$

where I is the indicator function. The generalisation error of a random forest is determined by

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \tag{3}$$

where $P_{X,Y}$ is the probability over the X, Y feature space (Breiman, 2001).

5.3. Artificial neural networks

Artificial neural networks, specifically multi-layer perceptrons (MLPs), have proved useful for research in emotion recognition from speech (Huber et al., 2000; Petrushin, 2000).

In the WEKA toolkit, ANNs are implemented as the multi-layer perceptron. Experiments with different network architectures led us to find highest accuracy using a one-hidden layer MLP with 38 input units, 60 hidden units, and two output units. An early stopping criteria based on a validation set consisting of 10% of the training set is used in all classification experiments involving the MLP. This ensures that the training process stops when the mean-squared error (MSE) begins to increase on the validation set and reduces overfitting (Haykin, 1999). The learning rate was set to 0.2 which is the default setting in WEKA.

5.4. K-nearest neighbours

K-nearest neighbours is another instance-based classification method introduced by Cover and Hart (1967). This algorithm has proved popular with vocal emotion recognition (Dellaert et al., 1996; Yacoub et al., 2003) due to its relative simplicity and performance comparable to other methods.

As with the K^* algorithm, the assumption for instance-based classifiers is that new instances will have the same class as pre-classified instances if they are close in feature space. For the *K*-nearest neighbour classifier, the nearest *K* neighbours of the current instance are retrieved (from some database of training instances) and the target class which the majority share is used as the class for the current instance (Cleary et al., 1995).

In our experiments, setting $K = 5$ performed best on the NATURAL data set. More information can be found in (Aha and Kibler, 1991).

5.5. K^* instance-based classifier

K^* is an instance-based learning algorithm based on the work of (Cleary et al., 1995). It uses a similarity function to classify test cases based on training cases which have a high similarity. In this way, it is much like the *K*-nearest neighbour method (described above), however, the distance measure used by K^* is based on entropy. Further detail on K^* can be found in the paper by Cleary et al. (1995).

5.6. Feature selection techniques

In order to optimise the classification time and accuracy, three feature selection algorithms were applied to each data set. The first was a step-wise forward selection, which is a well known technique for data reduction (Blum and Langley, 1997). Beginning with an initially empty set, a single feature is added at each step. Each unique feature set is tested with a subset evaluator. Each feature set is then ranked by classification accuracy and recorded. When the process is finished, the highest ranked feature set is retained.

The second feature selection algorithm employed was principal components analysis (PCA), another well known technique for data reduction and compression (Anton, 2000).

The third algorithm used was a genetic search which has been popular in recent research (Dieterle, 2003; Emmanouilidis et al., 1999; Vafaie and De Jong, 1992). A genetic search of the feature space mimicks biological evolution by "mutating" chromosomes (feature sets). Genes (individual features) make up the chromosomes which are initially randomly turned on or off (set to "0" = off or "1" = on).

Beginning with an initial population of randomly generated chromosomes, each chromosome is passed through a fitness function (for example, a classification model is generated and tested with the current chromosome) which ranks each member of the current generation according to its fitness (classification accuracy). Those chromosomes with the greatest fitness are "selected" and mated, with a mutation probability that introduces or removes one or more genes. When a stopping criteria has been met, such as a maximum number of generations, the process stops and ideally an optimal feature set is produced. A full description of genetic algorithms with examples can be found in (Goldberg, 1989).

6. Ensemble classification methods

Ensembles of classifiers generally combine several base classification schemes into a larger meta classifier. For ensemble classifiers to improve over the best performing base classifier, they must comprise accurate base classifiers. However, the base classifiers must also have high disagreement between one another in order to maintain diversity (Dietterich, 2002). For example, if a voting scheme is made up of several highly accurate base classifiers that cast the same prediction, then there is little improvement over simply using one of the base classifiers. The complexity involved in building the meta classifier must be outweighed by the improvement in classification accuracy.

6.1. Unweighted vote

The voting scheme we use is built by combining the aforementioned base classifiers: SVM with RBF kernel, random forest, K^* instance-based learner, KNN with $K = 5$, and multi-layer perceptron. Under this ensemble scheme, each classifier is trained with the same data. To measure performance, a test set is presented to each base classifier. The class predictions from each base classifier are then counted and the target class with the most votes is selected as the final prediction.

With unweighted voting, the class predictions of the base-level classifiers are summed and the class with the highest number of votes determines the prediction for the ensemble (Shipp and Kuncheva, 2002). For a voting ensemble with n classifiers, the output prediction (V_p) is determined by the following equation:

$$V_{\mathrm{p}} = \begin{cases} X & \text{when } \sum_{i=0}^{n} X_i > \sum_{j=0}^{n} Y_j, \\ Y & \text{when } \sum_{i=0}^{n} X_i < \sum_{j=0}^{n} Y_j, \\ \text{tie} & \text{when } \sum_{i=0}^{n} X_i = \sum_{j=0}^{n} Y_j, \end{cases} \tag{4}$$

where X and Y denote the predictions of the base classifiers for a two-class problem. In cases where an even number of base classifiers is used, there is potential for a tie when half of the classifiers vote for one class, and the other half vote for the opposition class. To avoid this problem, we use an odd number of base classifiers.

Because the confidence information contained in the prediction of each base level classifier is not taken into consideration, the resulting vote is unweighted, with all base level classifiers having equal input to the vote.

6.2. Stacked generalisation

Stacked generalisation, or stacking, is an approach to combining predictions from multiple classifiers. Introduced by (Wolpert, 1992), this method takes the predicted target classes of several different (or similar) base or level-0 classifiers and uses those to train a meta-learner or level-1 classifier. The meta-learner, typically a series (one for each target class) of linear models such as multi-response linear regression (MLR), uses the level-0 predictions and the target classes to determine which classifiers are correct or incorrect and generates a higher level prediction based on this.

a

Attrs	*Class*
$AttrVec_1$	*a*
$AttrVec_2$	*b*
$AttrVec_3$	*b*
$AttrVec_4$	*c*
⋮	⋮
$AttrVec_n$	*a*

b

a	*b*	*c*
0.90	0.05	0.05
0.15	0.70	0.15
0.10	0.80	0.10
0.20	0.20	0.60
⋮	⋮	⋮
0.80	0.10	0.10

c

$Classifier_1$			$Classifier_2$				$Classifier_N$			
a	*b*	*c*	*a*	*b*	*c*		*a*	*b*	*c*	class = *a*?
$P_{1,a1}$	$P_{1,b1}$	$P_{1,c1}$	$P_{2,a1}$	$P_{2,b1}$	$P_{2,c1}$	…	$P_{N,a1}$	$P_{N,b1}$	$P_{N,c1}$	1
$P_{1,a2}$	$P_{1,b2}$	$P_{1,c2}$	$P_{2,a2}$	$P_{2,b2}$	$P_{2,c2}$	…	$P_{N,a2}$	$P_{N,b2}$	$P_{N,c2}$	0
$P_{1,a3}$	$P_{1,b3}$	$P_{1,c3}$	$P_{2,a3}$	$P_{2,b3}$	$P_{2,c3}$	…	$P_{N,a3}$	$P_{N,b3}$	$P_{N,c3}$	0
$P_{1,a4}$	$P_{1,b4}$	$P_{1,c4}$	$P_{2,a4}$	$P_{2,b4}$	$P_{2,c4}$	…	$P_{N,a4}$	$P_{N,b4}$	$P_{N,c4}$	0
	⋮			⋮				⋮		⋮
$P_{1,an}$	$P_{1,bn}$	$P_{1,cn}$	$P_{2,an}$	$P_{2,bn}$	$P_{2,cn}$	…	$P_{N,an}$	$P_{N,bn}$	$P_{N,cn}$	1

d

$Classifier_1$	$Classifier_2$		$Classifier_N$	
a	*a*		*a*	class = *a*?
$P_{1,a1}$	$P_{2,a1}$	…	$P_{N,a1}$	1
$P_{1,a2}$	$P_{2,a2}$	…	$P_{N,a2}$	0
$P_{1,a3}$	$P_{2,a3}$	…	$P_{N,a3}$	0
$P_{1,a4}$	$P_{2,a4}$	…	$P_{N,a4}$	0
⋮	⋮		⋮	⋮
$P_{1,an}$	$P_{2,an}$	…	$P_{N,an}$	1

Fig. 3. Illustration of Stacking and StackingC on a three-class data set (*a*, *b*, *c*) with *n* training examples and *N* base classifiers. $P_{i,jk}$ denotes the class prediction from classifier *i* for class *j* on example *k* (from Seewald, 2002). (a) original training set (b) class probability distribution (c) meta training set for class *a*, Stacking with MLR and (d) meta training set for class *a*, StackingC with MLR.

StackingC, introduced by Seewald (2002), is an improvement over the original algorithm. It works by using only target predictions which are associated with the target class during training and testing. This has the effect of reducing the dimensionality of the meta-learning phase. The learning process is substantially faster by the encoding of meta-data in the level-1 feature space and uses prediction probabilities rather than actual target classes, which improves performance on multi-classed data. These prediction probabilities carry confidence information, which, when combined with a multi-response linear regression meta-learner, offers a modest improvement in classification accuracy for some multi-class problems (Seewald, 2002).

A comparison of the meta training sets used for Stacking and StackingC is shown in Fig. 3. Fig. 3a shows the original training set with the associated target class. Fig. 3b lists the probability distribution for each example. Fig. 3c shows the meta training set for Stacking. It can be seen that the training set includes the probabililities for each class whereas for StackingC, only the probabilities for the target class are used (see Fig. 3d).

In this study, StackingC is designed with the same five algorithms described in Section 5. The model is built by training each classifier individually on the same training set. The multi-response linear regression classifier is trained on the output predictions of the base classifiers. Performance is then measured by presenting the model with examples from the test set. The base classifiers output predictions which are then used by the MLR classifier to determine whether each base classifier will predict correctly.

7. Results and discussion

In this section we compare the classification accuracies on the NATURAL and ESMBS data sets using the classification methods introduced in the previous section. The methodology for acquiring the results below is as follows. Using the full NATURAL and ESMBS data sets, we perform feature selection using the three methods described in Section 5.6. First, the principal components are calculated for both data sets. Next, the other two feature selection methods, forward selection and genetic search, are employed using the SVM with RBF kernel as the subset evaluator. The choice is made to use the SVM with RBF kernel for the subset evaluator because it was the highest performer in the initial base classifier selection experiment and can be relied upon to describe the most relevant feature set. Next, we build the classifier ensembles and perform classification experiments on these using the original data sets as well as the subsets produced from feature selection.

For all classification experiments we employed 10 × 10-fold stratified cross-validations over the data sets. In other words, each classification model is trained on nine tenths of the total data and tested on the remaining tenth. This process is repeated ten times, each with a different partitioning seed, in order to account for variance between the partitions.

7.1. Performance of base classifiers

In Tables 6 and 7 we list the confusion matrices for the different base classifiers on the NATURAL and ESMBS data sets. It can be seen that for the NATURAL data set, the classification accuracy for neutral is much higher than that of anger. This is due to the unbalanced data (155 anger/233 neutral) in this set.

For the ESMBS data set, it is easily seen that anger and sadness are classified with high accuracy (generally >90%), where other emotions such as happiness and fear have much lower accuracies (sometimes <50%). This inter-class confusion is also common in human listeners (Scherer, 2003; Nwe, 2003). Emotion classes which oppose each other such as anger and sadness are much more easily separated than those classes with similar characteristics such as happiness and surprise.

For both data sets, the SVM with RBF kernel shows the highest performance. The random forest is the second best for the ESMBS data set, but is outperformed by the KNN on the NATURAL data set.

These results show that the ESMBS data set, while having six classes, is almost as accurately classified as the NATURAL data set, which only has two classes. Randomly classifying the ESMBS data set would show an average rate of about 16.67% whereas a randomly classifying the NATURAL data set would show an average rate of 50%.

This highlights the difficulties involved in using data collected from natural environments. The emotion represented is subtle and highly varied due to the uncontrolled nature of the method. Even two class problems such as this show quite low classification accuracies. Conversely, the

Table 6
Confusion matrices for the base classifiers on the NATURAL data set

	Anger	Neutral
(a) SVM (RBF)		
Anger	**67.94**	32.06
Neutral	17.08	**82.92**
Correctly classified (%): **76.93**		
(b) MLP		
Anger	**67.16**	32.84
Neutral	22.19	**77.81**
Correctly classified (%): **74.25**		
(c) KNN ($K = 5$)		
Anger	**64.26**	35.74
Neutral	16.44	**83.56**
Correctly classified (%): **75.85**		
(d) K^*		
Anger	**62.90**	37.10
Neutral	24.16	**75.84**
Correctly classified (%): **70.67**		
(e) RF		
Anger	**66.58**	33.42
Neutral	22.62	**77.38**
Correctly classified (%): **73.07**		

Table 7
Confusion matrices for the base classifiers on the ESMBS data set

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
(a) SVM (RBF)						
Anger	**91.24**	0.00	0.00	1.86	0.00	6.90
Disgust	0.08	**67.36**	15.27	11.78	1.63	3.88
Fear	0.23	16.67	**59.69**	13.64	4.81	4.96
Happiness	1.71	17.67	16.12	**49.61**	1.55	13.33
Sadness	0.00	3.10	4.03	1.24	**91.63**	0.00
Surprise	7.21	3.02	5.27	12.95	0.00	**71.55**
Correctly classified (%): **71.85**						
(b) MLP						
Anger	**90.54**	0.00	0.08	2.02	0.00	7.36
Disgust	0.00	**56.59**	20.31	14.88	2.48	5.74
Fear	0.31	19.30	**46.67**	16.12	8.45	9.15
Happiness	5.35	14.19	18.06	**42.40**	1.86	18.14
Sadness	0.00	3.41	4.81	1.63	**90.16**	0.00
Surprise	9.61	4.81	6.28	13.41	0.00	**65.89**
Correctly classified (%): **65.37**						
(c) KNN ($K = 5$)						
Anger	**93.41**	0.00	0.39	1.40	0.00	4.81
Disgust	0.93	**60.23**	20.00	12.40	1.47	4.96
Fear	0.08	23.33	**53.18**	12.17	4.81	6.43
Happiness	7.29	27.44	18.76	**31.78**	2.25	12.48
Sadness	0.00	7.21	9.84	2.87	**79.92**	0.16
Surprise	20.39	9.15	8.29	9.69	0.00	**52.48**
Correctly classified (%): **61.83**						
(d) K^*						
Anger	**91.32**	0.00	0.78	0.85	0.00	7.05
Disgust	0.70	**57.75**	20.47	11.63	4.57	4.88
Fear	1.55	25.81	**45.12**	16.36	3.64	7.52
Happiness	5.43	18.91	22.71	**37.21**	1.01	14.73
Sadness	0.00	5.04	14.57	3.64	**76.12**	0.62
Surprise	12.64	4.65	9.22	18.99	0.00	**54.50**
Correctly classified (%): **60.34**						
(e) RF						
Anger	**94.88**	0.00	0.54	0.39	0.00	4.19
Disgust	0.08	**65.81**	19.38	11.63	0.62	2.48
Fear	1.40	23.88	**47.60**	11.71	7.13	8.29
Happiness	5.12	22.79	18.68	**36.51**	0.54	16.36
Sadness	0.00	4.50	3.57	0.78	**91.01**	0.16
Surprise	6.98	7.60	4.65	12.40	0.00	**68.37**
Correctly classified (%): **67.36**						

results from the ESMBS data set show several important points. First, the classification accuracies are very similar to that from human listeners, as we saw in Section 2. Second, due to the high classification rates (which are much greater than chance, as mentioned above), we can see that the methods (features used, extraction methods, and classification algorithms) followed in this research are sound. Therefore, we can be confident in the results from the NATURAL data set.

7.2. Performance of ensemble classifiers

Next we present the performance statistics for the StackingC and vote ensembles. For stacked generalisation, the final prediction is based on a meta-classifier which is trained on the class probabilities and targets for each training example. Stacked generalisation attempts to predict

Table 8
Confusion matrices for the ensemble classifiers on the NATURAL data set

	Anger	Neutral
(a) StackingC		
Anger	**66.52**	33.48
Neutral	15.28	**84.72**
Correctly classified (%): **77.45**		
(b) Unweighted vote		
Anger	**69.03**	30.97
Neutral	15.97	**84.03**
Correctly classified (%): **78.04**		

Table 9
Confusion matrices for the ensemble classifiers on the ESMBS data set

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
(a) StackingC						
Anger	**94.81**	0.00	0.00	0.70	0.00	4.50
Disgust	0.08	**67.05**	16.98	10.08	2.25	3.57
Fear	0.08	19.15	**52.95**	10.70	7.75	9.38
Happiness	4.50	15.66	15.19	**44.73**	1.32	18.60
Sadness	0.00	1.94	2.64	0.70	**94.73**	0.00
Surprise	5.43	3.88	3.33	8.53	0.00	**78.84**
Correctly classified (%): **72.18**						
(b) Unweighted vote						
Anger	**95.89**	0.00	0.08	0.00	0.00	4.03
Disgust	0.39	**65.35**	17.75	10.93	0.62	4.96
Fear	0.31	20.93	**53.02**	12.64	4.96	8.14
Happiness	4.34	16.51	17.13	**45.12**	1.55	15.35
Sadness	0.00	1.78	3.95	1.16	**93.10**	0.00
Surprise	9.46	4.65	4.26	10.85	0.00	**70.78**
Correctly classified (%): **70.54**						

when the base classifiers will be incorrect. The unweighted vote simply sums the predictions of each class from the base classifiers and picks the most popular class.

The confusion matrices for the ensemble classifiers on the NATURAL data set are presented in Tables 8 and 9 shows the confusion matrices for the same ensembles on the ESMBS data set. On the NATURAL data set, both StackingC and the unweighted vote show improvement over the base classifiers in Table 6. Interestingly, anger is predicted more accurately for the vote while neutral speech is predicted less accurately.

For the ESMBS data set, StackingC performs better than all base classifiers, while the unweighted vote performs better than all but the support vector machine. Anger and sadness are both accurately classified, while the accuracies for happiness and fear are much lower. Anger shows a higher rate than that of StackingC, but sadness shows a lower classification rate. The vote is significantly less accurate for surprise than StackingC, with 70.78% and 78.84% respectively. Like the results on the base classifiers presented above, these predictions are in line with the accuracies for human classification.

7.3. Performance after feature selection

As mentioned above, because the SVM with RBF kernel is the most accurate, it is used for feature selection where a subset evaluator is required. A subset evaluator is required for the forward selection and the genetic search, since there must be a way of measuring the performance of the newly generated data set at each stage in the process. For principal components analysis, no subset evaluator is needed.

Feature selection is performed on both the NATURAL and ESMBS databases independently, meaning the process yields different feature subsets for each database. Table 10 shows the resulting feature subsets for each database.

Principal components analysis yields exactly the same feature sets for each database. To aid in the labelling of PCA selected features, we transform the principal components back into the original feature space and kept only the top 25 principal components.[2] Every attribute from the pitch and first formant frequency (F1) contour is retained. The majority of attributes from the F2 contour are retained, with the exception of the value at the last voiced segment and the range. No attributes for F3 are retained, hinting that this entire feature group may not add any variance to the data set. All energy attributes are kept, save the range and mean.

The feature sets resulting from forward selection do not seem to show much correlation between the NATURAL and ESMBS data sets. The F0 attributes are shared except that the value at the last voiced segment and the range are retained for the ESMBS data set.

The F1 features mean, minimum, maximum, and standard deviation are favoured for NATURAL, while maximum is discarded for ESMBS. F2 attributes compare similarly for each data set when compared with F1 attributes, except the value at the last voiced segment for F1/NATURAL is retained and the F2 standard deviation is discarded for ESMBS. F3 attributes are very different between data sets. They are sparsely retained for NATURAL but densely retained for ESMBS. This may be due to the subtlety of emotion in the NATURAL data set and the clear, concise nature of emotion in ESMBS due to the different data collection methods described in Section 2.

Mean energy is valued for both data sets, while the values at the first and last voiced segments seem important only for ESMBS. Forward selection on both data sets retain all of the rhythm-based statistics. Of interest is the fact that forward selection for the NATURAL data set resulted in 23 features retained, where for the ESMBS data set, 30 features are retained.

[2] For classification, the principal components are not transformed back into the original feature space.

For the search using the genetic algorithm, features for F0, F1, and F2 are almost identical with a high number of attributes retained. For F1, the maximum is discarded for ESMBS and for F2, the value at the first voiced segment is discarded for NATURAL but retained for ESMBS. It is the opposite case for the value at the last voiced segment: retained in NATURAL but discarded in ESMBS. The F3 attributes seem quite useful for both data sets in comparison to the other selection techniques, but more so for NATURAL. All energy features are kept for NATURAL, whereas all except the minimum and value at the first voiced segment for ESMBS. All features relating to rhythm are retained. Interestingly, the genetic search retains the highest number of features for both data sets, compared to forward selection.

Table 11 shows a summary of results with feature selection on both the NATURAL and ESMBS data sets. Figs. 4 and 5 show the results in graphical form for easier

Table 10
Resulting feature subsets after feature selection

No.	Description	Dataset					
		NATURAL			ESMBS		
		PCA	FW	GA	PCA	FW	GA
1	F0 mean	•	•	•	•	•	•
2	F0 minimum	•	•	•	•	•	•
3	F0 maximum	•	•	•	•	•	•
4	F0 standard deviation	•	•	•	•	•	•
5	First F0 value	•		•	•		•
6	Last F0 value	•		•	•	•	•
7	F0 range	•		•	•	•	•
8	F1 mean	•	•	•	•	•	•
9	F1 minimum	•	•		•	•	
10	F1 maximum	•	•	•	•		
11	F1 standard deviation	•	•	•	•	•	•
12	First F1 value	•		•	•	•	•
13	Last F1 value	•	•	•	•	•	•
14	F1 range	•		•	•	•	•
15	F2 mean	•	•	•	•	•	•
16	F2 minimum	•	•	•	•		•
17	F2 maximum	•	•	•	•	•	•
18	F2 standard deviation	•	•	•	•		•
19	First F2 value	•			•	•	•
20	Last F2 value			•		•	
21	F2 range			•		•	•
22	F3 mean			•		•	
23	F3 minimum		•	•		•	•
24	F3 maximum		•	•		•	
25	F3 standard deviation			•		•	•
26	First F3 value					•	•
27	Last F3 value		•	•		•	•
28	F3 range			•			•
29	Energy mean		•	•		•	•
30	Energy minimum	•	•	•	•		
31	Energy maximum	•	•	•	•	•	•
32	Energy standard deviation	•		•	•		•
33	First energy value	•		•	•	•	
34	Last energy value	•		•	•	•	•
35	Energy range		•	•			•
36	Speaking rate		•	•		•	•
37	Average length of unvoiced segments	•	•	•	•	•	•
38	Average length of voiced segments		•			•	•
Count		25	23	34	25	30	31

PCA = principal components analysis; FW = forward selection; GA = genetic algorithm. PCA data sets have been transformed back into the original feature space for labelling purposes and have the top 25 principal components retained.

Table 11
Average percentages of correctly classified instances from the NATURAL and ESMBS data sets for all classification methods

	Dataset							
	NATURAL				ESMBS			
	ORIG	PCA	FW	GA	ORIG	PCA	FW	GA
SVM (RBF)	76.93	75.98	**79.20**	75.95	71.85	63.94	70.72	**72.05**
MLP	74.25	72.06	75.15	73.99	65.37	61.12	66.71	66.86
*K**	70.67	66.55	71.19	71.68	60.34	44.32	58.13	61.43
RF	73.07	66.73	73.99	72.47	67.36	53.85	66.77	69.04
StackingC	77.45	75.49	**79.28**	77.73	72.18	63.59	72.44	**73.29**
Vote	78.04	75.57	**79.43**	77.83	70.54	59.97	69.38	**72.30**

For acronyms in the data set column, ORIG = original feature set; PCA = principal components analysis; FW = forward selection; GA = genetic algorithm.

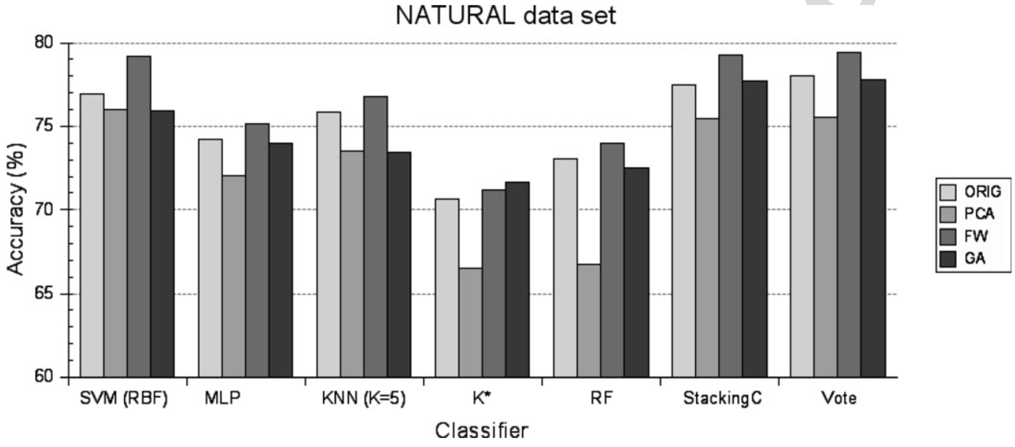

Fig. 4. Average percentages of correctly classified instances from the NATURAL data set for all classification methods. ORIG = original feature set; PCA = principal components analysis; FW = forward selection; GA = genetic algorithm. Unweighted vote combined with forward selection performs best.

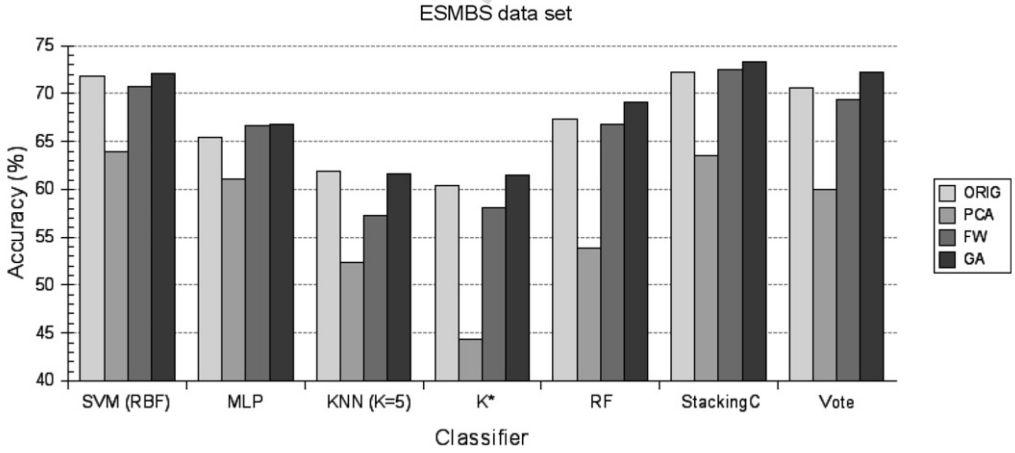

Fig. 5. Average percentages of correctly classified instances from the ESMBS data set for all classification methods. ORIG = original feature set; PCA = principal components analysis; FW = forward selection; GA = genetic algorithm. StackingC combined with the genetic search performs best.

comprehension. Forward selection proves to be the most accurate feature selection method for the NATURAL data set, proving more accurate for every classifier except *K**, where the genetic search yields a better feature set. The genetic search proves more accurate than forward selection and PCA on the ESMBS data set. PCA is always worse than even the original feature set, which is surprising, as PCA often has success for accurate feature reduction in other studies (Lee and Narayanan, xxx).

Observing the results more closely, we can see that forward selection on ESMBS actually worsens classification. This is likely due to the fact that the subset evaluation for the forward selection search is done using the SVM. Had each classifier evaluated its own intermediate subsets during selection, the resulting feature sets would likely have been better adapted to those classifiers.

With respect to the performance of the classifiers, there is a clear improvement with using the ensemble methods. For the NATURAL data set, we can see that the voting scheme and StackingC perform slightly better than the base classifiers using the original feature set, and marginally better than the SVM using the forward selection set. The improvement is more significant when we look at the results for the genetic search.

For the ESMBS data set, the ensemble methods perform significantly better on the original data sets compared to the base classifiers. Here, however, it is StackingC which shows the best performance by almost 2% over the voting scheme.

7.4. Summary of results

In summary, forward selection and the voting scheme performed best on the NATURAL data set, while the genetic search and StackingC performed best on the ESMBS data set. In general, the accuracies of between the different emotion classes remained constant over all classification methods.

In the ESMBS data set, anger and sadness were most accurately classified, followed by surprise, disgust, fear, and happiness. For the NATURAL data set, neutral speech was always classified more accurately than angry speech. The likely reason for this is that the data set was slightly unbalanced (60% neutral versus 40% angry). Overall, classification rates on the NATURAL data set were lower than expected. The main problem in using spontaneous emotional speech is the lack of control that the researcher has over the experiment. Lack of control can lead to unbalanced data, often having some background noise and levels of specific emotions that are difficult to quantify.

However, because automatic classification on the ESMBS data set showed success comparable to human listeners, we can be confident that the features utilised for describing the variation between emotional classes are very good.

The results also show us the inherent differences between the two data collection methods. Acted data may lead to inflated results, while data from real-world situations yields much lower classification rates, but paints a more realistic picture of applied automatic emotion recognition.

8. Conclusion and future work

In this paper we explored the performance of two ensemble speech classification schemes in comparison to several traditional base-level classifiers. Ensemble methods for classification have generally been overlooked for studies in emotion recognition. As seen in this paper, even simple methods such as combining predictions of base classifiers with a voting scheme can show a modest improvement in prediction accuracy. Further improvement could be gained by experimenting with different combinations of base-level classifiers.

This paper also explored the differences between portrayed and natural emotional speech. Portrayed speech yields the researcher a high amount of control over the emotion expressed, but fails to accurately model the subtle nature of real-world emotion. This leads to inflated classification accuracy when compared to natural emotion. Emotion collected from natural situations, on the other hand, offers the researcher virtually no control over the emotions expressed, and variance throughout the emotion classes is high. However, using portrayed emotional speech, while not necessarily useful in real-world situations, can provide a basis for investigating the acoustic differences between the different emotion classes.

We succeeded in building a speech database for spoken affect classification. The database is used in the framework for automatic emotion classification from speech. This framework is to be deployed in a call-centre where customers are interacting with human and/or machine representatives and will help with the management of customer disputes.

Future work includes processing more speech data from the call-centre environment which will be useful in determining recognition rates for a broader range of emotion. In addition, we hope to compare other methods of combining base-level classifiers. An important aspect relating to the application of this system is that it must constantly be updated as new speech data passes through it. Therefore, incremental learning and efficient retraining approaches will be considered as part of the ongoing research.

Acknowledgements

This study was funded by Technology for Industry Fellowships (TIF), New Zealand. The authors are grateful for the use of the speech database provided by Mabix International as well as the speech database provided by Tin Lay Nwe.

References

Aha, D., Kibler, D., 1991. Instance-based learning algorithms. Machine Learning 6, 37–66.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human–computer dialog. Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002). Denver, Colorado.

Anton, H., 2000. Elementary Linear Algebra. Von Hoffman Press.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E., 2003. How to find trouble in communication. Speech Communication 40, 117–143.

Blum, A., Langley, P., 1997. Selection of relevant features and examples in machine learning. Aritificial Intelligence 97, 245–271.

Breazeal, C., Aryananda, L., 2002. Recognition of affective communicative intent in robot-directed speech. Autonomous Robots 12, 83–104.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Cleary, J.G., Trigg, L.E., K*: An instance-based learner using and entropic distance measure. In: ICML, 1995, pp. 108–114.

Cover, T.T., Hart, P.E., 1967. Nearest neighbour pattern classification. IEEE Transactions on Information Theory 13, 21–27.

Cowan, M., xxx. Pitch and intensity characteristics of stage speech, Arch. Speech, Supplement to December Issue.

Davitz, J.R., 1964. Personality, perceptual, and cognitive correlates of emotional sensitivity. In: Davitz, J.R. (Ed.), The Communication of Emotional Meaning. McGraw-Hill, New York.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996), Philadelphia, PA, pp. 1970–1973.

Devillers, L., Vasilescu, I., Lamel, L., 2002. Annotation and detection of emotion in a task-oriented human–human dialog corpus. In: Proceedings of ISLE Workshop on Dialogue Tagging, Edinburgh.

Dieterle, F., 2003. Multianalyte quantifications by means of integration of artificial neural networks, genetic algorithms and chemometrics for time-resolved analytical data, Ph.D. thesis.

Dietterich, T.G., 2002. Ensemble learning. In: Arbib, M.A. (Ed.), The Handbook of Brain Theory and Neural Networks. The MIT Press, Cambridge, Massachusetts, pp. 405–408.

Elfenbein, H.A., Ambady, N., 2002. On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological Bulletin 128, 203–235.

Emmanouilidis, C., Hunter, A., MacIntyre, J., Cox, C., 1999. Multiple-criteria genetic algorithms for feature selection in neurofuzzy modeling. In: Proceedings of the International Joint Conference on Neural Networks, Washington, USA, pp. 4387–4392.

Fairbanks, G., Hoaglin, L.W., 1941. An experimental study of the durational characteristics of the voice during the expression of emotion. Speech Monograph 8, 85–91.

Fairbanks, G., Pronovost, W., 1939. An experimental study of the pitch characteristics of the voice during the expression of emotion. Speech Monograph 6, 87–104.

Fonagy, I., 1978. A new method of investigating the perception of prosodic features. Language and Speech 21, 34–49.

Fonagy, I., 1981. Emotions, voice and music. In: Sundberg, J. (Eds.), Research Aspects on Singing, Royal Swedish Academy of Music No. 33, pp. 51–79.

Fonagy, I., Magdics, K., 1963. Emotional patterns in intonation and music. Kommunikationsforsch 16, 293–326.

Frick, R.W., 1986. The prosodic expression of anger: differentiating thread and frustration. Aggressive Behaviour 12, 121–128.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading, Mass.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Prentice-Hall, Upper Saddle River, New Jersey.

Huber, R., Noth, E., Batliner, A., Buckow, J., Warnke, V., Niemann, H., 1998. You beep machine – emotion in automatic speech understanding systems. Proceedings of the Workshop on Text, Speech, and Dialog. Masark University, pp. 223–228.

Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., Niemann, H., 2000. Recognition of emotion in a realistic dialogue scenario. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000), Vol. 1, Beijing, China, pp. 665–668.

Johnson, W.F., Emde, R.N., Scherer, K.R., Klinnert, M.D., 1986. Recognition of emotion from vocal cues. Arch. Gen. Psych. 43, 280–283.

Lee, C.M., Narayanan, S., xxx. Towards detecting emotion in spoken dialogs 13 (2).

Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., 2004. Emotion recognition based on phoneme classes. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP 2004), Jeju Island, Korea.

Liscombe, J., Riccardi, G., Hakkani-Tnr, D., 2005. Using context to improve emotion detection in spoken dialog systems. Interspeech, 1845–1848.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., Stroeve, S., 2000. Approaching automatic recognition of emotion from voice: a rough benchmark. In: Proceedings of the ISCA Workshop on Speech and Emotion, Belfast, Northern Ireland, pp. 200–205.

Murray, I., Arnott, J., 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. Journal of the Acoustical Society of America 93, 1097–1108.

Nakatsu, R., Nicholson, J., Tosa, N., 1999. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In: Proceedings of the International Conference on Multimedia Computing and Systems, Florence, Italy.

Nwe, T.L., 2003. Analysis and detection of human emotion and stress from speech signals, Ph.D. thesis, Department of Electrical and Computer Engineering, National University of Singapore.

Nwe, T.L., Foo, S.W., De Silva, L.C., 2003. Speech emotion recognition using hidden markov models. Speech Communication 41, 603–623.

O'Shaughnessy, D., 2000. Speech Communications: Human and Machine, Second edition. IEEE Press, New York.

Oster, A., Risberg, A., 1986. The identification of the mood of a speaker by hearing impaired listeners. Speech Transmission Lab. – Q. Prog. Stat. Rep. 4, 79–90.

Petrushin, V., 2000. Emotion recognition in speech signal: experimental study, development, and application. In: Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China.

Platt, J., 1998. Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge, Massachusetts.

Polzin, T.S., Waibel, A., 2000. Emotion-sensitive human–computer interfaces. In: Proceedings of the ISCA Workshop on Speech and Emotion, Belfast, Northern Ireland.

Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, New Jersey.

Salovey, P., Kokkonen, M., Lopes, P., Mayer, J., 2004. Emotional intelligence: what do we know? In: Manstead, A.S.R., Frijda, N.H., Fischer, A.H. (Eds.), Feelings and Emotions: The Amsterdam Symposium. Cambridge University Press, Cambridge, UK, pp. 321–340.

Scherer, K.R., 1996. Adding the affective dimension: a new look in speech analysis and synthesis. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996), Philadelphia, PA.

Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Communication 40, 227–256.

Seewald, A.K., 2002. How to make stacking better and faster while also taking care of an unknown weakness. In: Proceedings of the 19th International Conference on Machine Learning, San Francisco, California.

Shipp, C.A., Kuncheva, L.I., 2002. Relationships between combination methods and measures of diversity in combining classifiers. Information Fusion 3, 135–148.

Skinner, E.R., 1935. A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness. Speech Monograph 2, 81–137.

Talkin, D., 1995. A robust algorithm for pitch tracking (rapt). In: Kleijn, W., Paliwal, K. (Eds.), Speech Coding and Synthesis. Elsevier Science B.V., The Netherlands, pp. 495–518.

Vafaie, H., De Jong, K., 1992. Genetic algorithms as a tool for feature selection in machine learning. In: Proceedings of the 4th International Conference on Tools with Artificial Intelligence, Arlington, VA.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, NY.

Williams, C.E., Stevens, K.N., 1972. Emotions and speech: some acoustical correlates. Nonverbal Communication: Readings with Commentary, second edition. Oxford University Press, New York.

Wolpert, D.H., 1992. Stacked generalization. Neural Networks 5, 241–260.

Yacoub, S., Simske, S., Lin, X., Burns, J., 2003. Recognition of emotion in interactive voice systems. In: Proceedings of Eurospeech 2003, 8th European Conference on Speech Communication and Technology, Geneva, Switzerland.