Available online at www.sciencedirect.com
ScienceDirect
Journal of Network and
Computer Applications 30 (2007) 1356–1365
Journal of NETWORK and COMPUTER APPLICATIONS
www.elsevier.com/locate/jnca

Voting ensembles for spoken affect classification

Donn Morrison[a], Liyanage C. De Silva[a,*]

[a]Institute of Information Sciences and Technology, Massey University, Palmerston North, Private bag 11222, New Zealand

Received 20 September 2006; accepted 20 September 2006

Abstract

Affect or emotion classification from speech has much to benefit from ensemble classification methods. In this paper we apply a simple voting mechanism to an ensemble of classifiers and attain a modest performance increase compared to the individual classifiers. A natural emotional speech database was compiled from 11 speakers. Listener-judges were used to validate the emotional content of the speech. Thirty-eight prosody-based features correlating characteristics of speech with emotional states were extracted from the data. A classifier ensemble was designed using a multi-layer perceptron, support vector machine, K^* instance-based learner, K-nearest neighbour, and random forest of decision trees. A simple voting scheme determined the most popular prediction. The accuracy of the ensemble is compared with the accuracies of the individual classifiers.

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

1. Introduction

Ensemble methods for classification have been gaining greater acceptance in many fields of applied machine learning. Improvements in computing power and memory allow more complex models to be generated and trained in less time.

The increase in human–computer interaction in recent years has led to a marked increase in research on emotion recognition and modelling. It is now desirable to design computer systems and robots that respond to the affective states of humans, enabling more natural

*Corresponding author.
E-mail address: l.desilva@massey.ac.nz (L.C. De Silva).

1084-8045/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jnca.2006.09.005

interaction. Such applications are useful in areas where humans interact with automated systems like call-centres, computer-aided learning, or interactive films (Picard, 1997).

In this paper, we focus on the application of spoken affect classification in a call-centre. For example, a customer telephones a customer service representative (CSR) and has concerns or questions about services, account information, bill payments, etc. In many cases, the customer is telephoning to resolve a dispute. During the call, emotion may be expressed by the customer and/or the CSR. It is helpful for the CSR to know when such a situation is arising and to take steps to ensure the customer remains satisfied with the service. Dissatisfied customers are likely to cause further problems or switch to a competitor, which in turn affects the potential profits of the company. Additionally, a team lead or manager may want to inquire on the status of any currently active calls in order to help coach new or inexperienced CSRs (see Fig. 1).

A call can be monitored for emotional variance by periodically making assessments on the speech signal. Features correlating vocal affect with emotional states are extracted from the speech signal using a variety of signal processing techniques. Statistics on these features are also calculated and converted into feature vectors which are then input to a classifier. The classifier, trained on similar data, assigns the vector to a class and the software displays the output on the CSRs terminal.

The rest of the paper is organised as follows. In Section 2 we present our methodology, including data collection, feature extraction, and classification methods. In Section 3, we introduce a simple machine learning technique to the field which has been overlooked in the past. Section 4 shows our experimental results comparing the voting scheme with each of the base classifiers individual performance. A discussion on these results follows. Finally, Section 5 gives our conclusions and avenues for future work.

2. Methodology

2.1. Data collection

The application of spoken affect recognition in the real-world has often suffered due to a lack of natural speech data (Batliner et al., 2003). There are difficulties collecting natural,

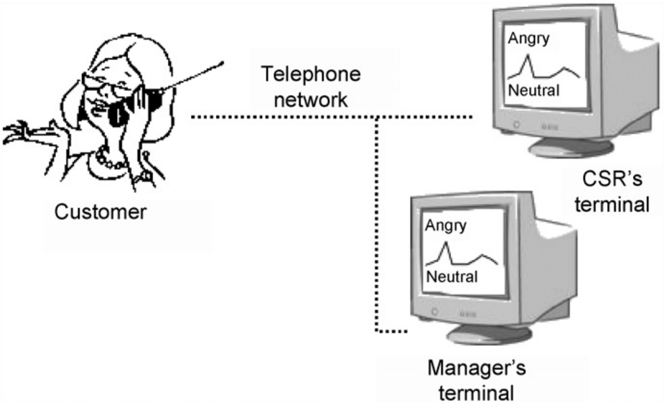

Fig. 1. Affect recognition in a call-centre. Real-time assessments can be made from the speech signal and displayed on the CSRs computer terminal, allowing improved response to customer emotion.

spontaneous emotional speech. It is unethical to record speakers without their consent, and the quality and content of the recordings are often poor and sparse. Because of this, most studies have used actors to gather emotional speech (Dellaert et al., 1996; Polzin and Waibel, 2000). However, databases of acted speech do not accurately reflect real-world spontaneous dialogue.

In spontaneous dialogue, affect is often subtly represented and difficult to detect, even for humans (Nwe et al., 2003). Other studies attempt to elicit more accurate content by inducing emotion in naïve speakers (Ang et al., 2002). In these situations, data are collected from participants interacting with systems designed to induce different emotions, for example, a malfunctioning appointment scheduling system that causes irritation and anger in the subjects (Huber et al., 2000). Although this technique brings the data closer to the real-world, the participants are not in a real scenario where stress can be accurately modelled.

In contrast to most past research, we collected real-world affective speech data from a call-centre providing customer support for an electricity company. Customers telephoned with general queries, problems and comments and each call was handled by a CSR. The distribution of affective content in the data is mainly neutral speech, with the second largest subset representing angry callers.

Due of the low distributions of happiness, sadness, surprise, disgust, and fear, 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. 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. (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.

After an initial manual segmentation and classification, the data set comprised 190 angry utterances and 201 neutral utterances totalling 391. However, to ensure that the manual classifications were objective, nine listener-judges were instructed to classify the entire data set. After the results of the listener-judges were available, 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 the NATURAL data set.

2.2. Acoustic correlates to anger

The fundamental frequency (F0) contour has been shown to vary depending on the emotional state being expressed. Early research discovered that neutral or unemotional speech has a much narrower pitch range than that of emotional speech (Cowan,). It was also 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). It was discovered in Williams and Stevens (1972) that 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 (Williams and Stevens, 1972). Frustration, which has similar but less extreme physiological effects as anger, has a higher fundamental frequency than neutral speech (Frick, 1986). Anger is described as having "an increase in mean pitch and mean intensity" (Scherer, 1996). Downward slopes are also noted on the pitch contour.

Fig. 2 shows two example pitch contours for angry and neutral utterances from the database. It can be seen that the angry sample shows the wider range, with downward slopes, and the neutral sample shows a much narrower and flatter contour.

The formant frequencies (F1, F2, F3) have also been noted to contain emotional markers. It was 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 (Williams and Stevens, 1972). 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 displayed a "uniform formant structure and glottal vibration patterns," contrasting the "irregular" formant contours of fear, sadness, and anger. Further, it has been found that angry speech has a noticeably increased energy envelope (Fonagy, 1981).

The speaking rate has been used in previous research (Dellaert et al., 1996; Petrushin, 2000; Ang et al., 2002). Fear, disgust, anger, and happiness often have a higher speaking rate, while surprise has a normal tempo and sadness a reduced articulation rate (Williams and Stevens, 1972). Anger has an increased speech rate, and "pauses forming 32% of total speaking time" (Fairbanks and Hoaglin, 1941; Fonagy, 1981).

2.3. 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*.

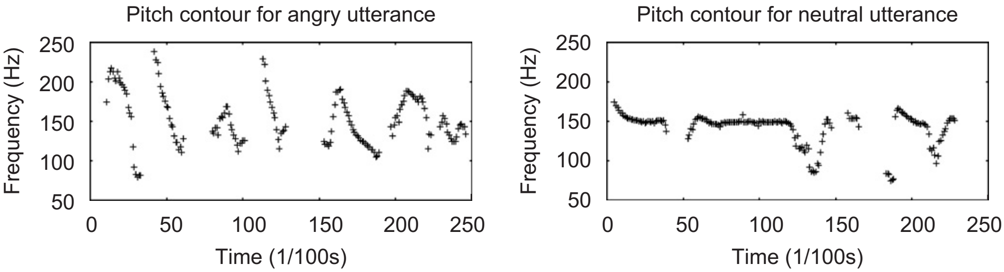

Fig. 2. Pitch contours for anger and neutral speech samples from the NATURAL speech corpus. The contour for the angry utterance has a much wider range and downward slopes are apparent. The contour for the neutral utterance has a narrow range which represents properties of calm speech.

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 1.

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 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 estimated by counting the number of syllables, which is roughly equal to the number of voiced-to-unvoiced and unvoiced-to-voiced transitions (hereafter referred to only as voiced–unvoiced transitions) during the utterance. A segment (one or more consecutive frames) 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.

Table 1
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.

2.4. Classification

Classification was performed using WEKA (Waikato environment for knowledge analysis). 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. Table 2 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 (MLP) (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.

2.4.1. 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. SVMs 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 RBF. As shown in Table 2, RBF performed better on our data set. The RBF kernel is defined as

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

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 $C = 8.0$.

2.4.2. Random forests

Random forests, invented by Breiman (2001), consist of ensembles of tree predictors. These tree ensembles are a method of growing a “forest” of decision trees by selecting

Table 2
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

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.

Breiman (2001) states 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).

2.4.3. *Artificial neural networks*

Artificial neural networks, specifically 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 MLP. 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.

2.4.4. *K^* instance-based classifier*

K^* is an instance-based learning algorithm based on the work of Cleary and Trigg (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 below), however, the distance measure used by K^* is based on entropy (Cleary and Trigg, 1995).

Further detail on K^* can be found in the paper by Cleary and Trigg (1995).

2.4.5. *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 and Trigg, 1995).

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

3. Ensemble with unweighted vote

In this paper, we will explore the advantages of using a simple unweighted voting scheme to create an ensemble from the five base-level classifiers. With unweighted voting, the predictions of the base-level classifiers are summed for each class 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:

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

where X and Y denote the predictions of the base classifiers. 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.

4. Experimental results

All classification experiments were conducted using 10 × 10-fold cross-validation. Cross-validation is a technique used to reduce variance in the results. The data set is divided into 10 equally sized subsets and at each iteration, one subset is held out and used for testing while the other nine subsets are used to train the models. This is repeated such that each subset is held out as a testing set. This process is then repeated ten times, each time using a different seed to generate the partitions.

Table 3 lists confusion matrices for the five base classifiers introduced in the previous section as well as the unweighted voting scheme. The SVM with RBF kernel shows the highest performance out of the base classifiers. However, combining the predictions of each base classifier using the unweighted voting scheme clearly increases classification accuracy. These results indicate that employing ensemble techniques on real-world data can lead to improved classifier generalisation.

5. Conclusion and future work

In this paper we explored the use of a simple unweighted voting scheme to combine the predictions of base-level classifiers. Our results show that there is a modest performance increase on the data set used. Further improvement could be gained by experimenting with different combinations of base-level classifiers. We also succeeded in building a

Table 3
Confusion matrices for the five base classifiers compared with the unweighted vote

		Anger	Neutral
(a)	SVM with RBF kernel		
	Anger	**67.94**	32.06
	Neutral	17.08	**82.92**
(b)	Random forest		
	Anger	**66.58**	33.42
	Neutral	22.62	**77.38**
(c)	One-hidden-layer perceptron		
	Anger	**67.16**	32.84
	Neutral	22.19	**77.81**
(d)	K^* instance-based learner		
	Anger	**62.90**	37.10
	Neutral	24.16	**75.84**
(e)	K-nearest neighbours (with $K = 5$)		
	Anger	**64.26**	35.74
	Neutral	16.44	**83.56**
(f)	Unweighted vote		
	Anger	**69.03**	30.97
	Neutral	15.97	**84.03**

(a) Instances correctly classified (%): **76.93**. (b) Instances correctly classified (%): **73.07**. (c) Instances correctly classified (%): **74.25**. (d) Instances correctly classified (%): **70.67**. (e) Instances correctly classified (%): **75.85**. (f) Instances correctly classified (%): **78.04**.

speaker-independent framework for spoken affect classification for use in a call-centre environment.

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.

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 such a 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.

References

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

Ang J, Dhillon R, Krupski A, Shriberg E, Stolcke A. Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: The International conference on spoken language processing (ICSLP 2002), Denver, CO, 2002.

Batliner A, Fischer K, Huber R, Spilker J, Noth E. How to find trouble in communication. Speech Communication 2003;40:117–43.

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

Cleary JG, Trigg LE. K: an instance-based learner using and entropic distance measure. In: ICML, 1995. p. 108–14.

Cover TT, Hart PE. Nearest neighbour pattern classification. IEEE Transactions on Information Theory 1967; 13:21–7.

Cowan M. Pitch and intensity characteristics of stage speech. Archive Speech 1: Supplementary to December Issue, pp. 1–92 (1936).

Dellaert F, Polzin T, Waibel A. Recognizing emotion in speech. In: The International conference on spoken language processing (ICSLP 1996), Philadelphia, PA, 1996. p. 1970–3.

Devillers L, Vasilescu I, Lamel L. Annotation and detection of emotion in a task-oriented human–human dialog corpus. In: ISLE workshop on dialogue tagging, Edinburgh, 2002.

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

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

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

Fonagy I, Emotions, voice and music. In: Sundberg J., editor. Research aspects on singing, Royal Swedish Academy of Music No. 33, 1981. p. 51–79.

Frick RW. The prosodic expression of anger: differentiating thread and frustration. Aggressive Behaviour 1986;12:121–8.

Haykin S. Neural networks: a comprehensive foundation. Upper Saddle River, NJ: Prentice-Hall; 1999.

Huber R, Batliner A, Buckow J, Noth E, Warnke V, Niemann H. Recognition of emotion in a realistic dialogue scenario. In: The International conference on spoken language processing (ICSLP 2000), vol. 1, Beijing, China, 2000. p. 665–8.

Murray I, Arnott J. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. Journal of the Acoustical Society America 1993;93:1097–108.

Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden Markov models. Speech Communication 2003;41:603–23.

Petrushin, V., 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, 2000.

Picard RW. Affective computing. Cambridge, MA: The MIT Press; 1997.

Platt J. Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf B, Burges C, Smola A, editors. Advances in kernel methods—support vector learning. Cambridge, MA: MIT Press; 1998.

Polzin TS, Waibel A. Emotion-sensitive human–computer interfaces. In: The ISCA workshop on speech and emotion, Belfast, Northern Ireland, 2000.

Rabiner LR, Schafer RW. Digital Processing of Speech Signals. Englewood Cliffs, NJ: Prentice-Hall; 1978.

Scherer KR. Adding the affective dimension: a new look in speech analysis and synthesis. In: The International conference on spoken language processing (ICSLP 1996), Philadelphia, PA, 1996.

Shipp CA, Kuncheva LI. Relationships between combination methods and measures of diversity in combining classifiers. Information Fusion 2002;3:135–48.

Talkin D. A robust algorithm for pitch tracking (RAPT). In: Kleijn W, Paliwal K, editors. Speech coding and synthesis. The Netherlands: Elsevier Science; 1995. p. 495–518.

Vapnik V. The nature of statistical learning theory. NY: Springer; 1995.

Williams, C.E., Stevens, K.N., Emotions and speech: some acoustical correlates. In: Nonverbal communication: readings with commentary. 2nd ed. New York: Oxford University Press; 1972.

Yacoub S, Simske S, Lin X, Burns J. Recognition of emotion in interactive voice systems. In: Eurospeech 2003, eighth European conference on speech communication and technology, Geneva, Switzerland, 2003.