Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Speech Communication 49 (2007) 201–212

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech

Mohammad Shami, Werner Verhelst *

Laboratory for Digital Speech and Audio Processing, Department of ETRO-DSSP, Vrije Universiteit Brussel, Interdisciplinary Institute for Broadband Technology, Pleinlaan 2, 1050 Brussels, Belgium

Received 16 April 2006; received in revised form 20 November 2006; accepted 18 January 2007

Abstract

In this study, the robustness of approaches to the automatic classification of emotions in speech is addressed. Among the many types of emotions that exist, two groups of emotions are considered, adult-to-adult acted vocal expressions of common types of emotions like happiness, sadness, and anger and adult-to-infant vocal expressions of affective intents also known as "motherese". Specifically, we estimate the generalization capability of two feature extraction approaches, the approach developed for Sony's robotic dog AIBO (AIBO) and the segment-based approach (SBA) of [Shami, M., Kamel, M., 2005. Segment-based approach to the recognition of emotions in speech. In: IEEE Conf. on Multimedia and Expo (ICME05), Amsterdam, The Netherlands]. Three machine learning approaches are considered, *K*-nearest neighbors (KNN), Support vector machines (SVM) and Ada-boosted decision trees and four emotional speech databases are employed, Kismet, BabyEars, Danish, and Berlin databases.

Single corpus experiments show that the considered feature extraction approaches AIBO and SBA are competitive on the four databases considered and that their performance is comparable with previously published results on the same databases. The best choice of machine learning algorithm seems to depend on the feature extraction approach considered.

Multi-corpus experiments are performed with the Kismet–BabyEars and the Danish–Berlin database pairs that contain parallel emotional classes. Automatic clustering of the emotional classes in the database pairs shows that the patterns behind the emotions in the Kismet–BabyEars pair are less database dependent than the patterns in the Danish–Berlin pair. In off-corpus testing the classifier is trained on one database of a pair and tested on the other. This provides little improvement over baseline classification. In integrated corpus testing, however, the classifier is machine learned on the merged databases and this gives promisingly robust classification results, which suggest that emotional corpora with parallel emotion classes recorded under different conditions can be used to construct a single classifier capable of distinguishing the emotions in the merged corpora. Such a classifier is more robust than a classifier learned on a single corpus as it can recognize more varied expressions of the same emotional classes. These findings suggest that the existing approaches for the classification of emotions in speech are efficient enough to handle larger amounts of training data without any reduction in classification accuracy.

Keywords: Emotion recognition; Analysis of intent; Vocal expressiveness; Speech processing

1. Introduction

Affective computing, which is currently a very active research field, aims at the automatic recognition and synthesis of emotions in speech, facial expressions, or any other biological communication channel (Picard, 1997). Within the field of affective computing, this paper addresses the problem of the automatic recognition of emotions in speech.

Emotion recognition technology is important because of its instantaneous applicability and usefulness in a world where the acceptability of automated systems is based on the user's perception of its intelligence and receptiveness. In fact, existing automatic speech recognition systems can

[*] Corresponding author. Tel.: +32 2 629 36 43; fax: +32 2 629 28 83. *E-mail address:* wverhels@etro.vub.ac.be (W. Verhelst).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.01.006

benefit from the extra information that emotion recognition can provide (Ten Bosch, 2003; Dusan and Rabiner, 2005). It would be useful to produce speech transcripts that not only contain the words said by different speakers, but also the speaker's state or emotion under which the words were said. In (Shriberg, 2005), the authors emphasize the importance of modeling non-linguistic information embedded in speech to better understand the properties of natural speech. Such understanding of natural speech is beneficial for the development of human-machine dialog systems (Shriberg, 2005).

Automatic recognition of emotions in speech aims at building classifiers (or models) for classifying emotions in unseen emotional speech. As opposed to rule based approaches, data-driven approaches to the classification of emotions in speech use supervised machine learning algorithms (such as neural networks or support vector machines, etc.) that are trained on patterns of speech prosody. Typically, statistical measures of speech pitch, short-time spectra and intensity contours are used as features of the expression of emotions in speech. These features are provided as input to a machine learning algorithm along with the known emotional labels of a training set of emotional utterances. The output of the supervised learning phase is a classifier capable of distinguishing between the different emotional classes it was trained with.

Previous studies have focused on a number of different aspects of the emotion recognition problem. Some studies focus on finding the most relevant acoustic features of emotions in speech as in (Nwe et al., 2003; Fernandez and Picard, 2005; Cichosz and Slot, 2005). Other studies search for the best machine learning algorithm to use in constructing the classifier as in (Oudeyer, 2003) or investigate different classifier architectures as in (Breazeal and Aryananda, 2002). Lately, research has shifted towards investigating the proper time scale to use when extracting features as in (Shami and Kamel, 2005; Katz et al., 1996). Although utterance level approaches are the most common (Schuller et al., 2005; Cichosz and Slot, 2005; Oudeyer, 2003), segment-based approaches are becoming more popular. Segment-based approaches try to model the shape of acoustic contours more closely as in (Katz et al., 1996; Schuller et al., 2003; Batliner et al., 2003, 2005; Rotaru and Litman, 2005). In all of the mentioned studies, a single speech corpus is used for training and testing a machine learned classifier. To our knowledge, multi-corpus emotion recognition using parallel emotional corpora has not been attempted.

In this study, we make a comparison between a segment-based approach used in (Shami and Kamel, 2005) and an utterance based approach used in (Oudeyer, 2003). Four emotional speech databases are used, Kismet, BabyEars, Danish, and Berlin, and a number of supervised machine learning algorithms are evaluated. Furthermore, we attempt at performing multi-corpus machine learning experiments where the classifier is machine learned and tested using labeled data from more than a single database. In short, our aims are twofold, an estimation of the accuracy of the different feature extraction and machine learning approaches in single corpus experiments and the assessment of the robustness of the approaches in multi-corpus experiments utilizing parallel emotional speech corpora.

In Section 2, we describe the segment based and the utterance based feature extraction approaches that were used in our work, and the classifier evaluation scheme employed. Section 3 describes the speech corpora used. The single corpus and multi-corpus classification experiments performed are discussed in Sections 4 and 5, respectively. Finally, the conclusions and suggestions for further work are in Section 6.

2. Investigated approaches to the automatic classification of emotions

2.1. Segment-based approach to the classification of emotions (SBA)

The feature set of the segment-based approach (SBA) is made up of a battery of 12 statistical measures of pitch, intensity, and spectral shape variation. As shown in Table 1, six measures are used to describe the pitch contour, three for the intensity contour, and three for rate of change. The feature extraction process thus consists of two steps: acoustic parameter contour (time series data) extraction and calculation of the statistical measures of the extracted contours. The time series data (contours) of pitch, intensity, and Mel Frequency Ceptral Components (MFCCs) are first extracted from the raw speech signal. Subsequently, the statistical measures are calculated from those contours.

For pitch extraction PRAAT (Boersma and Weenink, 1996) is employed. PRAAT uses a simple yet accurate pitch extraction algorithm that is based on an autocorrelation method described in the work reported in (Boersma, 1993). This pitch extraction algorithm is known to be robust and highly accurate and has been used in many speech processing studies including those aimed at the recognition of the affective content of voice such as the study in (Oudeyer, 2003; Shami and Kamel, 2005; Shami and Verhelst, 2006). As a by-product of the pitch extraction process, the utterance is segmented into a sequence of N voiced segments.

As the flowchart in Fig. 1 shows, the speech sample as a whole is first summarized using statistical measures of spec-

Table 1
Feature set in the segment-based approach (SBA)

Pitch	Intensity	Speech rate
• Variance • Slope • Mean • Range • Max • Sum of abs delta	• Variance • Mean • Max	• Sum of Absolute delta MFCC • Variance of Sum of absolute delta MFCC • Duration

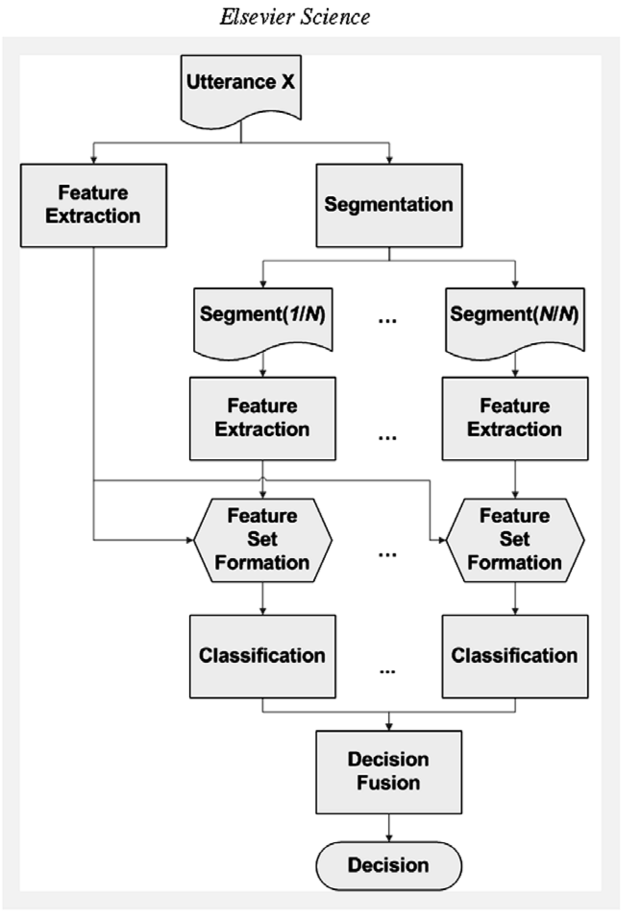

Fig. 1. Flowchart of the segment-based approach.

tral shape, intensity, and pitch contours as shown in Table 1. Specific to the SBA, however, is that the same battery of statistical measures that was calculated at the whole utterance level is recalculated for each of the N detected voiced segments. Now a feature vector consisting of both utterance level information and information local to the voiced segment is formed for each of the voiced segments. At classification time, and since class labels are provided for utterances as a whole, it is assumed that each of the voiced segments contains an expression of the emotion of the utterance it belongs to, and therefore it is given that same label. A segment classifier is then trained using these assumed segment labels and the feature vectors containing both local and global statistics. For the classification of whole utterances, the decisions made by the segment classifier for each of its voiced segments, expressed as a posteriori class probabilities, are aggregated to obtain a single utterance level classification decision. The utterance level classification decision is obtained by calculating the segment duration weighted sum of the segment a posteriori class probabilities as

$$P(C_n|F_{\mathrm{Utt}_A}) = \sum_{x=1}^{\mathrm{NumSegsInUtterance}} \mathrm{length}(\mathrm{Seg}_X) \times P(C_n|F_{\mathrm{Seg}_X}) \tag{1}$$

Using the utterance a posteriori class probabilities, and to make a final classification decision concerning the utterance, the Maximum A Posteriori rule (*MAP*) is used. An utterance represented by F_{Utt_X} is classified as C_w if:

$$P(C_{\mathrm{w}}|F_{\mathrm{Utt}_X}) = \arg\max\{P(C_n|F_{\mathrm{Utt}_X})\}, 1 < n < T \tag{2}$$

More details on the specifics of the SBA algorithm are in (Shami and Kamel, 2005).

2.2. Utterance based classification of emotions: the AIBO approach

The AIBO approach is a bottom up approach that relies on using an extensive feature set of low level statistics of prosodic parameters. The utterance is viewed as a single object and low level statistics of pitch, intensity and Delta MFCC are calculated and used in conjunction with a machine learning algorithm to construct a classifier from labeled data.

First the pitch, intensity, lowpass intensity, highpass intensity, and the norm of the absolute vector derivative of the first 10 MFCC components are extracted from the speech signal. Next, out of each of the five time series data four series are further extracted: the series of minima, the series of maxima, the series of the durations between local extrema of the 10 Hz smoothed curve, and the series itself. The last step is to calculate statistics on the extracted $5 \times 4 = 20$ series. Ten statistics are calculated on each of the series as shown in Table 2.

More details on the algorithm are in (Oudeyer, 2003).

2.3. Experimental setup

By combining the above feature sets AIBO and SBA with the machine learning algorithms considered, i.e., support vector machines (SVM), *K*-nearest neighbors (KNN), and Ada-boosted C4.5, six different emotion classifiers can be trained on a given speech corpus.

To estimate the performance of a classifier, the original speech corpus is split into two disjoint sets, a training set

Table 2
Feature set in the AIBO approach

Acoustic features	Derived series	Statistics
• Intensity • Lowpass intensity • Highpass intensity • Pitch • Norm of absolute vector derivative of the first 10 MFCC components	• Minima • Maxima • Durations between local extrema • The feature series itself	• Mean • Maximum • Minimum • Range • Variance • Median • First quartile • Third quartile • Inter-quartile range • Mean absolute value of the local derivative

and a test set. The training set contains the data used to generate the classifier. The testing set, which is used to measure the performance of a particular classifier, should be different from the data in the training set; otherwise, misleadingly overoptimistic results would ensue.

Unless otherwise specified, the corpus splitting scheme used is stratified *n*-fold cross validation. In *n*-fold-cross-validation, the labeled corpus S is randomly split into n disjoint subsets. Assuming that S contains K instances, then each of the disjoint subsets would contain K/n instances. Next, n classifiers are generated using the same learning algorithm and conditions but every time one out of the n subsets is left out of the training set and used as a testing set. Therefore, the remaining $(n - 1)$ subsets are used for training the classifier. The process is repeated n times, every time using a different subset for testing. Overall performance is then taken as the average of the performance achieved in the n runs. The most common value of n used in the literature is 10; therefore, this value is adopted. Additionally, we use stratified cross-validation where the n subsets contain approximately the same proportions of classes as the original dataset.

3. Speech corpora

The speech corpora used in this experiment are two infant directed corpora, Kismet and BabyEars and two adult directed corpora, Danish and Berlin. Both infant directed corpora contain expressions of non-linguistic communication (affective intent) conveyed by a parent to a preverbal child.

3.1. Kismet

The first corpus is a superset of the Kismet speech corpus that has been initially used in (Breazeal and Aryananda, 2002). The corpus used in this work contains a total of 1002 American English utterances of varying linguistic content produced by three female speakers in five classes of affective communicative intents. The classes are Approval, Attention, Prohibition Weak, Soothing, and Neutral utterances. The affective intents sound acted and are generally expressed rather strongly. Recording is performed with 16-bit per samples with occurrences of 8 and 22 kHz sampled recordings and under varying amounts of noise. The speech recordings are of variable length, mostly in the range of 1.8–3.25 s.

3.2. BabyEars

The second speech corpus is the BabyEars speech corpus that has been used in previous studies (Slaney and McRoberts, 2003; Shami and Kamel, 2005). The corpus consists of recordings in American English of six mothers and six fathers as they addressed their infants while naturally interacting with them. The emotions expressed in the speech recordings sound natural and unexaggerated. Three emotional classes are included in the corpus, namely: Approval, Attention, and Prohibition. The total number of recordings in the corpus is 509 recordings with unbalanced classes. The utterances are typically between 0.53 and 8.9 s in length. The breakdown of the emotional class distribution in the Kismet and the BabyEars databases are in Table 3.

Table 3
Emotional classes in the Kismet and BabyEars database pairs

Kismet		BabyEars	
Approval	185	Approval	212
Attention	166	Attention	149
Prohibition	188	Prohibition	148
Soothing	143		
Neutral	320		

Since the Kismet corpus contains two extra emotional classes that are not available in the BabyEars corpus, it was necessary to remove those two classes of emotions when performing multi-corpus experiments. Assuming that the recorders of the two corpora intended to have the same color of emotion under the same emotional label, the removal of the extra classes makes the two corpora compatible for machine learning experiments.

3.3. Berlin database

The emotional speech database in German was recorded at the Technical University of Berlin to study the acoustical features of emotional expression (Paeschke and Sendlmeier, 2000). Five female and five male actors uttered ten sentences in German that have little emotional content textually. Recordings were made using high-quality recording equipment in an anechoic chamber. Other data sources were recorded besides voice namely electro-glottograms and narrow transcripts. The total number of utterances is 493 divided among seven emotional classes: neutral, anger, fear, joy, sadness, disgust, and boredom. Recordings were made with 16-bit precision and at a sampling rate of 22 kHz.

3.4. Danish database

The Danish database is described in detail in (Engberg and Hansen, 1996). It consists of a combination of short and long utterances in Danish spoken by two male and two female speakers in five emotions. These emotions are neutral, surprised, happy, sad, and angry. Recordings were

Table 4
Emotional classes in the Berlin and Danish database pairs

Berlin		Danish	
Anger	127	Angry	52
Sadness	52	Sad	52
Happiness	64	Happy	51
Neutral	78	Neutral	133
Fear	55	Surprised	52
Boredom	79		
Disgust	38		

made with 16-bit precision and at a sampling rate of 20 kHz in a recording studio.

Similarly to the Kismet and BabyEars database pair, the Berlin and Danish databases share a number of emotional classes (Table 4). Only these common emotional classes were used in multi-corpus experiments.

4. Single corpus classification experiments

In this section, we present the outcome of the classification experiments performed on each of the four databases individually. We compare the two approaches considered, the segment-based approach and the AIBO approach, in addition to comparing the performance of three machine learning algorithms: support vector machines (SVM), *K*-nearest neighbors (KNN), and Ada-boosted C4.5.

4.1. Classification outcome

In all classification experiments, we use the implementations of machine learning algorithms in the data mining toolkit Weka (Witten and Frank, 2000) and Milk (Frank and Xu, 2003). In all the machine learning experiments whose results are summarized in Table 5:

- All accuracies are 10-fold cross validation percentage accuracies unless otherwise specified.
- Results reported as previous results on the same database in Table 6 are based on classification and evaluation schemes comparable to the experiments performed here except for those studies that are marked with an asterisk (*).
- To learn about the most commonly confused emotion classes, the confusion matrix resulting from using the AIBO approach and an SVM classifier are given in Appendix A.
- Whenever human perception accuracy tests exist for a database, the results are given along with machine learning classification accuracy.
- Reported baseline accuracy is the result of classifying all test utterances as the most common emotional class in the database.
- The parameters of the machine learning algorithms used are the default parameters provided by the software toolkits Weka and Milk.

4.2. Analysis of classification results

When comparing the approaches AIBO and the segment-based approach (SBA), one can see that they are competitive. When using the best performing machine learning algorithms, the AIBO approach gives better performance on the Berlin and Danish databases while the SBA is superior on the Kismet and BabyEars databases.

It is noteworthy to mention that there are similarities between the Kismet and the BabyEars databases on one hand, and the Berlin and Danish databases on the other. The major difference between the two database pairs is the type of emotions expressed. The Kismet/BabyEars pair contains affective communicative intents whereas the Berlin/Danish pair contains expressions of emotions like happiness, sadness, etc. The other difference between the two database pairs is the length of the emotional utterances, and consequently, the average number of voiced segments per utterance. Fig. 2 shows how the Berlin and Danish databases contain more voiced segments per utterance than the Kismet and BabyEars databases. These similarities and differences can explain why the AIBO approach seems to be more suitable for Berlin and Danish whereas the SBA works better with Kismet and BabyEars databases.

Table 5
Percentage classification accuracy in single corpus experiments

MLA	Kismet		BabyEars		Berlin		Danish	
	AIBO	SBA	AIBO	SBA	AIBO	SBA	AIBO	SBA
SVM	83.7	83.2	65.8	67.9	75.5	65.5	63.5	56.8
KNN	82.2	86.6	61.5	68.7	67.7	59.0	49.7	55.6
ADA-C4.5	84.63	81	61.5	63.4	74.6	46.0	64.1	59.7

Table 6
Previous results on the same databases

Kismet		BabyEars		Berlin		Danish	
Study	Acc.	Study	Acc.	Study	Acc.	Study	Acc.
Breazeal and Aryananda (2002)	81.9	Slaney and McRoberts (2003)	67	Paeschke and Sendlmeier (2000)	85	Ververidis and Kotropolos (2004)	50.6
		Shami and Kamel (2005)	77	Baseline	34	Hammal et al. (2005)	54
Shami and Kamel (2005)	89	Baseline	41.7			Ververidis and Kotropolos (2004)	67
Baseline	31.9					Baseline	50.8

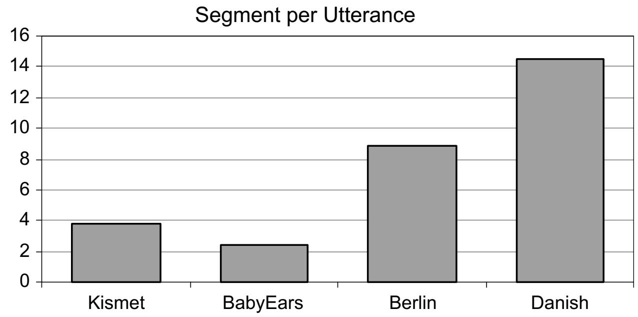

Fig. 2. Average number of voiced segments per utterance in the four databases.

4.3. *Effects of the machine learning algorithms chosen*

We now examine the effect of the machine learning algorithm (MLA) choice on classification performance. Performance scores can be found in Table 5 and are summarized in Table 7. Table 7 highlights for each MLA and for each specific database whether the best results where obtained with either the AIBO or the SBA approach. For example, using SVM, better performance is seen on both the Berlin and the Danish databases with the AIBO approach. Blank entries in Table 7 represent the cases where no clear winner is apparent. We can notice the following:

- SVM performs better on 2 out of 4 databases with the AIBO approach. This can be explained by the fact that SVM is immune to the detrimental effects of large and possibly noisy feature sets.
- KNN performs better on 3 out of 4 databases with the SB approach. The success of KNN with the SB approach can be related to the compactness of the feature set and the relevance of each of its member features.
- ADA-C4.5. ADA-Boosted Decision Trees is an MLA that builds linear combinations of C4.5 decision trees. ADA-C4.5 performs better with the AIBO approach on 2 out of 4 databases. Similar to SVM, ADA-C4.5 is known to provide automatic feature selection on large feature set. This allows the MLA to harness the full power of the large feature set without getting affected by irrelevant or redundant features.

4.4. *Comparison with previous results*

The experimental setup used in other studies that work with the same databases might differ from those used in these experiments in terms of the training/testing split and performance measure. Therefore, direct comparison is not completely feasible. Yet, if one ignores such mismatches in experimental setups we would find that the classification accuracies obtained in the experiments here are very competitive with previously published results. On the Kismet database best results using the AIBO approach with ADA-C4.5 yields 84.6% accuracy whereas best accuracy of 86.6% is obtained with the SBA and KNN. These accuracy rates compare well to previous results reported in (Breazeal and Aryananda, 2002). The authors of (Breazeal and Aryananda, 2002) employ mostly utterance level statistics of pitch and intensity and perform feature selection as they construct serial hierarchical classifiers using Guassian Mixture Models. This approach gives 81.9% accuracy on a subset of the Kismet database used here. More recent results on the Kismet database are reported in (Shami and Kamel, 2005). The SBA without feature selection gives 86.6% and with feature selection gives 89% in (Shami and Kamel, 2005).

Table 7
Comparison of MLA effectiveness across AIBO and SBA

	SVM	KNN	ADA-C4.5
Kismet	–	SBA	–
BabyEars	–	SBA	–
Berlin	AIBO	AIBO	AIBO
Danish	AIBO	SBA	AIBO

Previous results on the BabyEars database are in (Slaney and McRoberts, 2003; Shami and Kamel, 2005). In (Slaney and McRoberts, 2003) the authors use statistical features of spectral shape, pitch, and intensity contours with Gaussian Mixture Model classifiers and wrapper-based feature selection. Bootstrap accuracy[1] (as opposed to cross-validation accuracy) in (Slaney and McRoberts, 2003) is 67%. In (Shami and Kamel, 2005), wrapper-based feature selection and the SBA give a bootstrap accuracy of 77%. Best results achieved in the experiments reported here give 65.8% and 68.7% accuracy with AIBO and SBA respectively.

On the Berlin database, the AIBO and SB approaches give accuracies of 75.5% and 65.5% respectively. Both accuracies are lower than human classification accuracy of 85%. This suggests that there is definitely further room for improvement on the Berlin database.

On the fourth database, the Danish database, the AIBO and SB approaches give 64.1% and 59.7% accuracies respectively. In comparison, the approach in (Ververidis and Kotropolos, 2004) gives 50.7% accuracy when using Bayesian learning with Guassian PDFs. Hammal et al. (2005) obtains 54% accuracy with SVM classifiers that use Radial Basis Function Kernels. Finally, listening exper-

[1] In the '0.632' bootstrapping splitting scheme the training set is the same size as the original set. The training data is generated by resampling the original data with replacement k times, where k is the size of the original set. The testing data would consist of all the instances in the original dataset that were never selected in the training set. Performance is measured as error true according to:

$$e_{\text{true}} = 0.368 e_{\text{apparent}} + 0.632 \frac{1}{m} \sum_{i=1}^{m} e_{\text{boot}\, i}$$

where the apparent error is the one resulting from testing the classifier generated using the original set as both training and testing set. Bootstrap error is the error resulting from evaluating the classifier using a training bootstrap as a training set and a testing bootstrap as the testing set.

iments performed by the creators of the database gave 67% accuracy.

4.5. Comparison with human listeners

Additionally, the makers of the BabyEars database have performed a human listening test to see how accurately the emotions can be perceived. With the help of several raters the following was observed:

- 7 out of 7 raters unanimously labeled the utterance with the correct label 79% of the time.
- 5 out of 7 raters labeled the utterance with the correct label 85% of the time.
- 4 out of 7 raters labeled the utterance with the correct label 100% of the time.

On the Danish database, a listening experiment was performed with the help of 20 listeners (Engberg and Hansen, 1996). Overall correct listening rate is 67%. It was observed that listeners adapted to the speaking style of the actors during the listening sessions. The adaptation to the speaker enabled the listeners to raise their recognition accuracy by a 10% margin between the first 20 utterances and the last 20 utterances. The significant improvement in recognition accuracy underlines the impact of speaker dependence of automatic emotion recognition systems.

Further comparisons of the misclassification tendencies between automatic classification and human perception can be made from the confusion matrices of the experiments shown in Tables 8 and 9, respectively. Two class pairs are confused in human listening experiments: *Surprise*-*Happiness* and *Neutral*-*Sadness*. On the other hand, automatic classification with AIBO using SVM classification exhibits both *Neutral*-*Sadness* and *Surprise*-*Happiness* confusion and further confuses *Angry* with *Neutral*, and *Angry* with *Happy* emotions. It is very interesting to see similar emotion misclassification tendencies between automatic recognizers and human raters.

Table 8
Automatic classification confusion matrix of the Danish database

A (%)	B (%)	C (%)	D (%)	E (%)	←	Classified as
82.7	0	3.0	9.8	4.5	A	Neutral
5.8	50.0	32.7	0	11.5	B	Surprised
13.7	27.5	45.1	3.9	9.8	C	Happy
25.0	0	3.8	69.2	1.9	D	Sad
21.2	17.3	19.2	1.9	40.4	E	Angry

Table 9
Human listening confusion matrix of the Danish database

A (%)	B (%)	C (%)	D (%)	E (%)	←	Classified as
60.8	2.6	0.1	31.7	4.8	A	Neutral
10.0	59.1	28.7	1.0	1.3	B	Surprised
8.3	29.8	56.4	1.7	3.8	C	Happy
12.6	1.8	0.1	85.2	0.3	D	Sad
10.2	8.5	4.5	1.7	75.1	E	Angry

5. Multi-corpus classification experiments

As described in Section 3, the four databases in this study can be grouped in two pairs: the Kismet–BabyEars pair contains infant directed affective speech, while the Berlin–Danish pair contains adult directed emotional speech. In multi-corpus experiments and for each of the emotional database pairs, the following is performed. First, only the emotional classes that are common to both databases in consideration are kept in both databases. The remaining classes are removed.

Three kinds of experiments are performed on the paired databases.

(A) With-in corpus classification on each of the two databases is performed for comparative purposes.
(B) In off-corpus classification, a classifier is first machine learned using one corpus and, subsequently, tested on emotional samples from the other corpus.
(C) Integrated-corpus testing involves merging the two corpora into one speech corpus and then performing within-corpus testing on the resulting corpus. The goal is to examine the extent by which a classifier trained on the patterns of the similar emotions in both databases can distinguish among those patterns.

5.1. The BabyEars–Kismet database pair

Based on the results of Table 5, the SBA approach with KNN classifier has been found to be more suitable for the Kismet and BabyEars databases and was thus used here for classifying the three classes of affective speech that are common to the Kismet and BabyEars databases, namely *Approval*, *Attention* and *Prohibition*.

5.1.1. With-in corpus results

Table 10 shows the classification accuracy resulting from within-corpus classification experiments on both corpora. One can notice that a better classification rate is obtained on the Kismet corpus than on the BabyEars corpus. The difference can be attributed to two main causes. First, speaker variability in the BabyEars is higher than that in the Kismet corpus because it has more speakers of both genders in contrast to the Kismet corpus. Secondly, the type of speech found in the Kismet corpus contains stronger and more exaggerated expressions of the emotions.

Table 10
With-in corpus classification accuracy

Corpus	Classification accuracy (%)
BabyEars	65.40
Kismet	88.30

This makes the classification of the intended emotions in Kismet an easier task.

Reported accuracies in Table 10 differ from those in Table 5 even though the same experimental setup (in terms of classifiers and training/testing split) is used. This difference in results is due to the extra emotions that were removed in order to make the databases similar in the number of emotional classes used.

5.1.2. *Off corpus results*

The classification results in off-corpus experiments are shown in Table 11. Note that the baseline classification accuracy is calculated as the accuracy resulting from a classifier that always classifies test samples as belonging to the most frequent class in the test database.

From the results summarized in Table 11, it is possible to notice the following. First, when testing on the Kismet corpus and on the BabyEars corpus, the resulting accuracy is higher than the baseline accuracy. This suggests that the learned classifier has captured enough information about the emotional class found in the testing set by learning from the training set even though the two sets come from two different domains and are recorded under different conditions.

Training using the more varied BabyEars database and testing on Kismet database is found to be more successful than the other way around. This might be due to the fact that it is more difficult for a classifier trained on a corpus with female speakers only (Kismet corpus) to correctly classify samples from a speech corpus that has both male and female speakers (BabyEars corpus).

5.1.3. *Integrated corpus results*

In integrated-corpus experiments a total of 3 emotional classes are present in this setup. Overall classification accuracy obtained when the corresponding classes in the two corpora are merged is 74.6% correct. For comparison purposes, the resulting accuracy is plotted in Fig. 3 next to the results obtained in within corpus tests performed in the previous section.

The resulting classification accuracy in integrated corpus mode is somewhere in between the accuracies generated in within corpus testing. This might suggest that the patterns behind the different emotions in the two corpora do not overlap in the feature space. As an example, the “Approval” class of the Kismet corpus is not getting confused with, let us say, the “Attention” class of the BabyEars corpus and so forth. In order to examine in more detail the confusion tendencies in integrated corpus mode, the corresponding emotional classes of the two corpora were kept distinct. Therefore, there were six emotional classes in total, namely: {Approval, Attention Prohibition}_Kismet, and {Approval, Attention, Prohibition}_ BabyEars.

Table 11
Off-corpus classification results

Training set	Testing set	Classification accuracy (%)	Baseline accuracy (%)
BabyEars	Kismet	54.40	34.90
Kismet	BabyEars	45.00	41.70

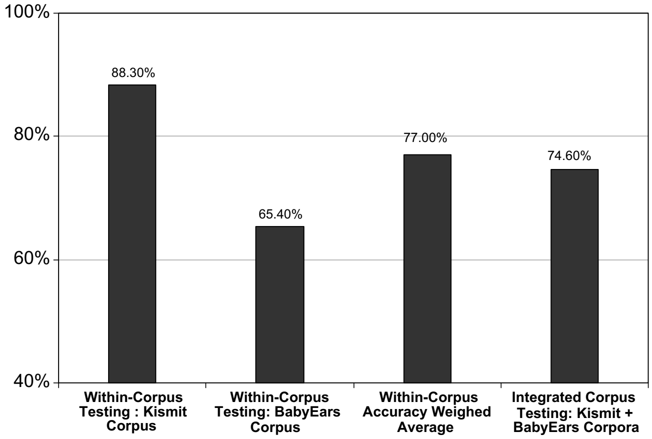

Fig. 3. Comparison of the classification accuracies obtained in different settings.

The confusion matrix resulting from machine learning and classification with these six classes is shown in Table 12. Cells in bold denote test utterances that have been correctly classified. One can see that Kismet samples are almost never confused with BabyEars samples. On the other hand, a significant number of BabyEars samples are confused with Kismet samples. The majority of those confused samples are actually classified correctly in terms of the conveyed emotion, which is very encouraging (32 samples with correct emotional label out of 50 corpus-confused samples, or 64% correct).

It is interesting to notice that similar classification accuracies are obtained when the emotional classes from the two databases are grouped before classification and when the classes are kept distinct as in Table 12. The resulting accuracy of six-way classification in Table 12 is 75.9% as compared to an accuracy of 74.6% in three-way classification.

In order to further analyse the proximity of the emotional classes in the above classification experiment we perform the following experiment. We apply clustering on the six-class integrated corpus of the Kismet/BabyEars databases using the *K*-means clustering algorithm imple-

Table 12
Confusion matrix in the case of integrated-corpus classification across six classes

a	b	c	d	e	f	←	Class
150	26	8	**1**	0	0	a	Ap_K
26	**140**	0	0	**0**	0	b	At_K
1	0	**186**	0	1	**0**	c	Pr_K
20	7	6	**105**	32	40	d	Ap_B
0	**2**	3	24	**86**	32	e	At_B
1	0	**9**	29	16	**93**	f	Pr_B

mented in Weka. We choose to cluster the instances into 6 clusters and we note the classes to clusters evaluation in Table 13.

Table 13 shows the classes to clusters evaluations taken as per class percentages (each row adds up to 100%). Analyzing the clustering results provides more insight into the distribution of the instances from all emotional classes in the feature space.

The Kismet and the BabyEars databases lie on sets of clusters with significant overlap. Specifically, the Kismet database is on clusters {A, B, C, D, E} whereas the BabyEars database is on clusters {A, C, D, E, F}. Furthermore, similar emotions in the two databases frequently lie on the same clusters. For example, *Kismet_approval* is on clusters {B, D, E} whereas *BabyEars_approval* falls on {D, E, F}. Similarly, *Kismet_prohibition* and *BabyEars_prohibition* lie on cluster {A}.

Furthermore, different emotions in the two databases lie on different clusters. In other words, the same cluster does not contain different emotions from different databases. For example, clusters {C, D, E} carry {*attention, approval, approval*} from both databases. Cluster F carries *approval* and *attention* from the BabyEars database only.

5.2. The Berlin–Danish database pair

In this section the AIBO approach is used on the four common emotional classes of the Danish and Berlin databases: neutral, happy, sad, and angry. The two classifiers SVM and KNN are employed.

5.2.1. With-in corpus results

Within corpus experiments show that for the same number of classes, a higher accuracy is obtained on the Berlin database in comparison with the Danish database (Table 14).

5.2.2. Off corpus results

In off-corpus testing, one corpus is used to build the classifier and the second is used for testing the generated classifier. The classification accuracies are shown in Table 15.

Table 15 shows that the obtained off-corpus classification accuracies are similar to baseline classification accuracies when either database is used for testing and using either SVM or KNN as the MLA.

Table 13
Classes to clusters evaluation in BabyEars Kismet integrated corpus

A	B	C	D	E	F	← Assigned to cluster
0	30.9	0	30.3	38.6	0.1	kismet_approval
0	0	99.8	0.2	0	0	kismet_attention
100.0	0	0	0	0	0	kismet_prohibition
0	3.2	0	19.0	31.0	46.7	babyears_approval
4.1	0	46.6	0	0	49.3	babyears_attention
98.3	0	0	0.7	1.0	0	babyears_prohibition

Table 14
Within corpus results using the AIBO approach and SVM

Database	Classification accuracy (%)
Danish	64.90
Berlin	80.7

Table 15
Off corpus classification results

Training	Testing	MLA	Classif. acc. (%)	Baseline acc. (%)
Berlin	Danish	SVM	20.8	24.3
Berlin	Danish	KNN	22.9	24.3
Danish	Berlin	SVM	52.6	46.2
Danish	Berlin	KNN	38.9	46.2

Off Corpus experiments on the Danish/Berlin databases (using the AIBO approach) show that generalization was not possible across databases. Similar to previously obtained results on the Kismet/BabyEars databases (using the segment-based approach), it seems that training on one database and testing on another database that shares common emotional classes is not possible in general with the existing approaches. The only case where there was some meaningful generalization across databases was for testing on Kismet with a classifier that was trained using the more varied BabyEars database, as discussed in Section 5.1.2 and even then the result stayed significantly below the within-database accuracies of both databases.

5.2.3. Integrated corpus results

In integrated corpus experiments the two speech corpora are merged into a single corpus. The resulting corpus is then randomly split into training and a testing set using 10-fold cross validation. The resulting classification accuracies obtained when the corresponding classes in the two corpora are combined, are shown in Table 16.

In order to examine in more detail the confusion tendencies in integrated corpus mode the following is performed. The corresponding emotional classes of the two corpora are not merged into a single corpus but are kept distinct. Therefore, there were 8 emotional classes in total, namely Danish_and Berlin_{neutral, happy, angry, sad}.

It can be observed in Table 17 that the instances belonging to one database are not being confused for instances belonging to the second database. Specifically, two different classes from the two databases are not getting confused for each other. This means that the different classes lie at different locations in the feature space.

Table 16
Classification results in integrated corpus tests

MLA	Classif. acc. (%)
SVM	72.2
KNN	66.83

Table 17
Confusion matrix in integrated corpus mode using Danish/Berlin database

a	b	c	d	e	f	g	h	←	Class
74	1	2	1	**0**	0	0	0	a	berlin_neutral
3	**36**	0	25	0	**0**	0	0	b	berlin_happy
4	0	**48**	0	0	0	**0**	0	c	berlin_sadness
1	25	0	**101**	0	0	0	**0**	d	berlin_anger
0	0	0	0	**106**	2	17	8	e	danish_neutral
0	**0**	0	0	7	**29**	2	13	f	danish_happy
0	0	**0**	0	21	4	**27**	0	g	danish_sad
0	0	0	**0**	11	16	2	**23**	h	danish_angry

Similar to Section 5.1.3, we apply clustering on the eight-class integrated corpus of the Danish/Berlin databases using the *K*-means clustering algorithm in order to examine the locations of the emotional classes in the feature space. We choose to cluster the instances into 8 clusters and we note the classes to clusters evaluation in Table 18.

Table 18 shows the classes to clusters evaluations taken as per class percentages (each row adds up to 100%). Clustering results show that the Danish and the Berlin databases lie on sets of clusters with little overlap. Specifically, the Danish is on clusters {A, B, D, E, G} whereas the Berlin database is on clusters {C, F, H}. Consequently, similar emotions in different databases lie on different clusters. For example, *berlin_happy* is on clusters {F, H} whereas *danish_happy* falls on {A, B, D, E, G}.

Furthermore, the expression of emotional classes in the Berlin database is less varied and more consistent than the expression of the same emotional classes in the Danish database. This point is supported by the observation that each of the Berlin emotional classes is represented by fewer clusters than each of the Danish classes. Table 18 shows that the cluster that has the most instance assignments of each of the four emotions in the Berlin database carries a higher percentage of the emotional class than is the case for the Danish database. The emotions {*neutral*, *happy*, *sadness*, *anger*} are assigned to a single cluster with a rate of {100%, 71.9%, 78.8%, 80.3%} in the Berlin database as compared to {33.8%, 45.1%, 53.8%, 48.1%} in the Danish database. The higher consistency of emotional expressions in the Berlin database explains the higher classification accuracy obtained in within corpus testing on the Berlin database as reported in Table 14.

Table 18
Classes to clusters evaluation in Danish/Berlin integrated corpus

A	B	C	D	E	F	G	H	←	Assigned to cluster
0	0	0	0	0	0	0	100.0		berlin_neutral
0	0	0	0	0	71.9	0	28.1		berlin_happy
0	0	78.8	0	0	0	0	21.2		berlin_sadness
0	0	0	0	0	80.3	0	19.7		berlin_anger
15.0	18.8	0	15.8	33.8	0	16.5	0		danish_neutral
3.9	7.8	0	45.1	7.8	2.0	33.3	0		danish_happy
7.7	7.7	0	26.9	3.8	0	53.8	0		danish_sad
7.7	5.8	0	48.1	9.6	0	28.8	0		danish_angry

5.2.4. Analysis of results

Off corpus testing on the two corpora as reported in Section 5.2.2 shows only little improvement over baseline classification. On the other hand, when the two corpora are merged into a single large corpus, classification accuracy is only slightly reduced compared to the single corpus experiments. An examination of the results of automatic clustering also provides evidence suggesting that emotional corpora of the same emotion classes recorded under different conditions can be used to construct a single classifier capable of distinguishing the emotions in the merged corpora. This is due to the fact that different emotions from different databases are being assigned to different clusters. The classifier learned using the merged corpora is more robust than a classifier learned on a single corpus because it can deal with emotions in speech that is recorded in more than one setting.

Automatic clustering of the emotional classes in the integrated corpora shows that the patterns behind the emotions in the Kismet/BabyEars integrated corpus are less database dependent than the patterns in the Berlin/Danish databases. Among the many possible explanations, one can argue that the emotional pattern could be linked to the language in which the emotions are being expressed. Since the Kismet and BabyEars databases are both American English corpora, the pattern of expression of the emotions is more generalizable across the two corpora. Another possible explanation for the higher pattern generalization in the Kismet/BabyEars corpora is in the nature of the emotions expressed in those corpora. As opposed to the Berlin/Danish databases, the Kismet/BabyEars corpora contain infant directed communicative intents. Infant directed communicative intents are generally regarded as culture and language independent (Fernald, 1992).

It is a well-known fact in machine learning that the more specific and uniform the training corpus is, the more accurate the classifier learned using that corpus. On the other hand, when the classifier is learned using a more heterogeneous corpus, the expected classification accuracy is usually less when the learned classifier is used to classify new instances. In our case, it turned out that using a heterogeneous emotional corpus (Kismet/BabyEars and Berlin/Danish database pairs) for constructing the classifier did not result in a notable deterioration in classification accuracy. In other words, the added robustness is not costly in terms of recognition accuracy.

6. Conclusion

Single corpus classification shows that the considered approaches AIBO and SBA are competitive. Specifically, the AIBO approach outperforms the SBA on the Berlin and Danish databases whereas the SBA gives better classification accuracy on the Kismet and BabyEars databases. The difference in performance reported on the same databases between the AIBO and the SBA is significant. The AIBO approach seems to be better suited for classification of emotions in emotion databases with long utterances whereas the SBA works better with short utterances.

The choice of the most effective machine learning algorithm (MLA) seems to depend on the approach (AIBO vs SBA). An approach that uses a large feature set of low level statistics such as the AIBO approach seems to work best with an SVM or an ADA-C4.5 classifier. Both machine learning algorithms are known to be successful in building accurate classifiers using large and possibly noisy feature sets which is the case in the AIBO approach. KNN performs best with the SBA, which is based on a more compact feature set than the AIBO approach.

Finally, the classification accuracies obtained in this study are very competitive with previously published results on the same databases. In fact, such competitive results are demonstrated with both classification approaches AIBO and SBA as shown in Section 4.

The results of this multi-corpus study underline the importance of performing more sophisticated performance measurements when evaluating supervised machine learning approaches to the classification of emotions in speech. In fact, off-corpus testing on both corpus pairs of parallel emotional classes reveals that there is little generalization happening for the same emotional classes across databases. Fortunately, when the two emotional corpora that share the same emotional classes are merged into a single large corpus, stratified cross-validation classification accuracy on the resulting database is only slightly reduced compared to the single database accuracies.

This suggests that emotional corpora with parallel emotion classes recorded under different conditions can be used to construct a single classifier capable of distinguishing the emotions in the merged corpora. The classifier learned using the merged corpora is more robust than a classifier learned on a single corpus because it can recognize more varied expressions of the same emotional patterns. Such findings suggest that the existing approaches for the classification of emotions in speech are efficient enough to handle larger amounts of training data without any reduction in classification accuracy. This way, more recordings expressing the same emotions in slightly different domains can continuously be added to the training corpus to produce a more robust classifier for the target emotions

Automatic clustering of the emotional classes in the integrated corpora shows that the patterns behind the emotions in the Kismet/BabyEars integrated corpora are less database dependent than the patterns in the Berlin/Danish databases. The Danish and the Berlin databases lie on sets of clusters with little overlap, consequently, similar emotions in the two databases also lie on distinct sets of clusters. If an ideal feature space could be employed, similar emotions belonging to different databases should be assigned to the same clusters.

To achieve the desired robustness, an alternative method to the use of more training data perhaps lies in a different direction. Adding robustness to the feature set that is used to represent the emotion in the utterance can compensate for the lack of vast amounts of training data. The design of better features and new paradigms is essential to the development of robust classification systems. It would be interesting to investigate the use of acoustic features that mimic the process of perception of emotions by people. Speech recognition research is already moving in the direction of using non-linguistic information from the speech signal to improve speech recognition accuracy (Dusan and Rabiner, 2005). For emotion recognition, the integration of knowledge from domains such as psychoacoustics is one step towards building emotion recognition systems that mimic human emotion perception.

Acknowledgements

The authors would like to thank the owners of the four emotional speech databases for providing their databases. The authors also thank the anonymous reviewers for their detailed and valuable comments. Parts of the research reported on in this paper were performed in the context of the IBBT projects Architectures for Mobile Community Content Creation (A4MC3) and Virtual Individual Networks (VIN), as well as in support of the HOA8 project Development of a research methodology and instrument for studying the quality of early parent-child relations at the Vrije Universiteit Brussel and the project Anty that enjoys the support of the Brussels region.

Appendix A. Classification confusion matrices

See Tables 19–22.

Table 19
Confusion matrix of the Kismet database

a	b	c	d	e		←	Classified as
144	25	6	6	4	|	a	Approval
27	137	0	0	2	|	b	Attention
3	0	166	1	18	|	c	Prohibition
6	1	2	127	7	|	d	Soothing
8	4	38	5	265	|	e	Neutral

Table 20
Confusion matrix of the BabyEars database

a	b	c		←	Classified as
158	27	27	|	a	Approval
34	91	24	|	b	Attention
37	25	86	|	c	Prohibition

Table 21
Confusion matrix of the Berlin database

a	b	c	d	e	f	g	←	Classified as
101	0	1	3	21	0	1	|	Anger
0	67	1	0	2	3	6	|	Boredom
3	2	26	1	1	2	3	|	Disgust
6	1	1	37	2	2	6	|	Fear
28	1	1	1	31	0	2	|	happiness
0	3	3	0	1	43	2	|	Sadness
0	4	1	3	1	2	67	|	Neutral

Table 22
Confusion matrix of the Danish database

a (%)	b (%)	c (%)	d (%)	e (%)	←	Classified as
82.7	0	3.0	9.8	4.5	|	Neutral
5.8	50.0	32.7	0	11.5	|	Surprised
13.7	27.5	45.1	3.9	9.8	|	Happy
25.0	0	3.8	69.2	1.9	|	Sad
21.2	17.3	19.2	1.9	40.4	|	Angry

References

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

Batliner, A., Steidl, S., Hacker, C., Nöth, E., Niemann, H., 2005. Tales of tuning – prototyping for automatic classification of emotional user states. In: Interspeech 2005, pp. 489–492.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proc. Institute of Phonetic Sciences of the University of Amsterdam, Vol. 17. pp. 97–110.

Boersma, P., Weenink, D., 1996. PRAAT: a system for doing phonetics by computer. Report of the Institute for Phonetic Sciences of the University of Amsterdam, Vol. 132. <http://www.praat.org>.

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

Cichosz, J., Slot, K., 2005. Low-dimensional feature space derivation for emotion recognition. In: Interspeech 2005, Lisbon, Portugal, pp. 477–480.

Dusan, S., Rabiner, L., 2005. On Integrating insights from human speech perception into automatic speech recognition. In: Interspeech 2005, Lisbon, Portugal.

Engberg, I.S., Hansen, A.V., 1996. Documentation of the Danish emotional speech database (DES). Internal AAU Report, Center for Person Kommunikation, Denmark.

Fernald, A., 1992. Human maternal vocalizations to infants as biologically relevant signals: an evolutionary perspective. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford.

Fernandez, R., Picard, R.W., 2005. Classical and novel discriminant features for affect recognition from speech. In: Interspeech 2005, Lisbon, Portugal, pp. 473–476.

Frank, E., Xu, X., 2003. Applying propositional learning algorithms to multi-instance data. Working Paper, Department of Computer Science, University of Waikato. <www.cs.waukato.nz/ml/milk>.

Hammal, Z., Bozkurt, B., Couvreur, L., Unay, D., Caplier, A., Dutoit, T., 2005. Passive versus active: vocal classification system. In: Proc. Eusipco 2005, Antalya, Turkey.

Katz, G., Cohn, J., Moore, C., 1996. A combination of vocal F0 dynamic and summary features discriminates between pragmatic categories of infant-directed speech. Child Dev. 67, 205–217.

Nwe, T., Foo, S., De Silva, L., 2003. Speech emotion recognition using hidden Markov models. Speech Comm. 41-4, 603–623.

Oudeyer, P., 2003. The production and recognition of emotions in speech: features and algorithms. Internat. J. Hum.–Comput. Stud. 59, 157–183.

Paeschke, A., Sendlmeier, W., 2000. Prosodic characteristics of emotional speech: measurements of fundamental frequency movements. In: Proc. ISCA ITRW on Speech and Emotion. Belfast, pp. 75–80.

Picard, R., 1997. Affective Computing. The MIT Press.

Rotaru, M., Litman, D., 2005. Using word-level pitch features to better predict student emotions during spoken tutoring dialogues. In: Interspeech 2005.

Schuller, B., Rigoll, G., Lang M., 2003. Hidden Markov model-based speech emotion recognition. In: IEEE Conf. on Multimedia and Expo (ICME04), Vol. I. Baltimore, USA, pp. 401–404.

Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., Rigoll, G., 2005. Speaker Independent speech emotion recognition by ensemble classification. In: IEEE Conf. on Multimedia and Expo (ICME05), pp. 864–867.

Shami, M., Kamel, M., 2005. Segment-based approach to the recognition of emotions in speech. In: IEEE Conf. on Multimedia and Expo (ICME05), Amsterdam, The Netherlands.

Shami, M., Verhelst, W., 2006. Automatic classification of emotions in speech using multi-corpora approaches. In: Proc. Second Annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2006), Antwerp, Belgium.

Shriberg, E., 2005. Spontaneous speech: how people really talk and why engineers should care. In: Eurospeech 2005, Lisbon, Portugal.

Slaney, M., McRoberts, G., 2003. A recognition system for affective vocalization. Speech Comm. 39, 367–384.

Ten Bosch, L., 2003. Emotions, speech and the ASR framework. Speech Comm. 40 (1–2), 213–225.

Ververidis, D., Kotropolos, C., 2004. Automatic speech classification to five emotional states based on gender information. In: Proc. Eusipco 2004, Vienna, Austria, pp. 341–344.

Witten, I.H., Frank, E., 2000. Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco.