BJÖRN SCHULLER / MARTIN WÖLLMER / FLORIAN EYBEN /
GERHARD RIGOLL

Prosodic, Spectral or Voice Quality? Feature Type Relevance for the Discrimination of Emotion Pairs[1]

As for every automatic pattern recognition task, the performance of emotion recognition is strongly dependent on the suitability of the used features. Therefore, determining optimal features provides the basis for reliable emotion classification. Prosodic features such as energy or pitch are widely used for emotion recognition since they are highly correlated to emotion. Together with voice quality, spectral features, and statistic functionals, they are extracted to form high dimensional feature vectors which often contain redundant information. Consequently, it is necessary to select the features which best reflect emotion. However, the question which features to use for what emotion or speaker group has not been answered so far. To get an insight into the relevance of specific feature types for different emotion discrimination tasks, three different databases of English emotional speech, which contain different types of emotion, were used for recognition experiments in this chapter: the 'Big Six' emotions as named by Ekman (1972), different stress-levels, and different interest-levels. A large set of 1,406 acoustic features was used. The features were grouped into three different categories: prosodic, spectral, and voice quality features. The performance of an emotion recogniser using exclusively certain feature groups or types (such as duration, energy, pitch, formants, cepstral coefficients, jitter, shimmer, or harmonics-to-noise ratio) was evaluated with and without subsequent automatic feature selection. Thereby 24 different combinations of

[1] The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

emotion pairs were considered and experiments were conducted speaker-independently. The overall goal was to answer the question which features best signal which emotion type. Further, redundancy would be reduced in order to operate in a lower dimensional feature space and save computational power while simultaneously improving recognition performance.

1. Introduction

Opposing related speech recognition tasks, the predominant question of optimal features is still an open issue for recognition of affect (McGilloway *et al.* (2000: 207-212), Cowie *et al.* (2001: 32-80), Pantic and Rothkrantz (2003: 1370-1390), Schuller *et al.* (2007b: 2253-2256)). It is well known that prosodic features based on speech rate, loudness, and pitch, as well as voice quality or articulatory information are highly correlated to emotion. Yet, it is not fully researched and determined which attributes best signal affect in general (Batliner *et al.* (2006: 240-245), Schuller *et al.* (2007b: 2253-2256)), let alone specific emotions, or speaker groups. In order to provide additional research and input to answer this yet unsolved issue, effects of features on emotion recognition performance have been investigated on multiple databases of English emotional speech.

Three sets were chosen to overcome singular effects arising from typical use of single sets. The sets covered a broad variety of emotions and situations reaching from elicited 'Big Six' emotions as named by Ekman (by the eNTERFACE set) to real-life stress-levels (by the SUSAS set), and to levels of interest (by the AVIC set) in spontaneous human conversational speech. Conditions throughout experiments were kept constant by the use of one systematically and fully automatically extracted large feature set. As a starting point, 37 acoustic Low-Level-Descriptors, which are well known to carry information about paralinguistic effects, were chosen. Following the typical static classification strategy, 19 statistical functionals were employed for each Low-Level-Descriptor. Hence, the obtained

multivariate time series of variable length was projected onto a single 1,406 dimensional feature vector as used among others in Batliner *et al.* (2008: 4497-4500), Schuller *et al.* (2008), and Vlasenko *et al.* (2007: 139-147). The functionals included the first four statistical moments, quartiles, extremes, ranges, positions, and zero-crossings. The obtained features were grouped into the following types: prosodic (duration, energy, and pitch), spectral (formants and cepstral), and voice quality (jitter, shimmer, and harmonics-to-noise ratio). The results were reported for the performance of each feature group individually, with respect to the aforementioned specific emotion pairs. The relevance of single features was revealed by additional employment of popular closed-loop floating search, and correlation analysis. To cope with current challenges and prepare for the future generation of emotion recognition, all experiments have been carried out speaker-independently.

The knowledge obtained can be used to design recognition engines more efficiently due to lower extraction effort and compact feature space representation. Additionally, sets tailored for emotion pairs can boost recognition performance by layer-wise optimised two-class decisions as in Support-Vector-Tree construction or Boosting.

The rest of this chapter is structured as follows: first, the emotional speech corpora are introduced in Section 2, following the acoustic features and their grouping in Section 3. In Section 4 the classification and selection of features is explained in detail and in Section 5 experimental findings with respect to feature type relevance are presented. A final conclusion is drawn in Section 6.

2. Emotional speech corpora

In this section, the three databases used for the investigation of feature type relevance with respect to the discrimination of emotion pairs are introduced. All speech samples in the databases are spoken in English. The sets contain Ekman's six basic emotions, stress levels, and interest in spoken language. They were chosen to cover a broad range

of emotions and user states. Additionally, weight has been laid on spontaneous emotions, which is in particular true for two of the three databases. Finally, large and popular sets were preferred.

2.1. Basic Emotions – The eNTERFACE Corpus

The eNTERFACE corpus is a public audio-visual emotion database introduced by Martin *et al.* (2000: 207-212). Emotions covered are Ekman's basic emotions (1972: 207-283), often referred to as the 'Big Six' and standardised in MPEG-4. 44 subjects from 14 nations are contained in the database. It consists of studio recordings of pre-defined spoken content whereas each subject was instructed to listen to six successive short stories, each of them eliciting a particular emotion. Further, they had to react to each of the situations and two experts judged whether the reaction expressed the emotion in an unambiguous way. Only then was the sample added to the database. The audio sample rate is 48 kHz; the bit depth is 16 bit. Overall, the database consists of 1,170 samples distributed among emotions as shown in Table 1.

Anger	*Disgust*	*Fear*	*Joy*	*Sadness*	*Surprise*	*Sum*
200	189	189	205	195	192	1170

Table 1. Distribution of speaker-turns among emotions in the eNTERFACE database.

2.2. Stress – The SUSAS Corpus

Next, the Speech Under Simulated and Actual Stress (SUSAS) audio database by Hansen *et al.* (1997: 1743-1746) was considered for the experiments presented here. These are spontaneous recordings in field noise. Herein, the 3,593 stress speech samples were used as shown in Table 2. These were recorded in fear and stress tasks in a helicopter and rollercoaster environment. Seven speakers, three of them female, in free fall stress situations are contained. Two different stress conditions were collected: medium and high stress. Further samples cover neutrality, fear, and screaming as classes. However, neutrality is too sparsely represented to be included. The SUSAS samples are

constrained to a 35 words vocabulary of short aircraft communication commands. All files are sampled in 8 kHz at 16 bit. The recordings are partly overlaid with heavy noise and background over-talk. This resembles realistic acoustic recording conditions as given in many application scenarios.

High Stress	*Med. Stress*	*Neutrality*	*Fear*	*Sum*
1202	1276	701	414	3593

Table 2. Distribution of speaker-turns among emotions in the SUSAS database.

2.3. Interest – The AVIC Corpus

Finally, the evaluation of feature type relevance is provided on the Audiovisual Interest Corpus (AVIC) covering interest types as further classes (Schuller *et al.* (2007a: 3037-3041)). In the recording scenario setup an experimenter and a subject were sitting on opposite sides of a desk. The experimenter played the role of a product presenter and led the subject through a commercial presentation. The subject's role was to listen to explanations and topic presentations of the experimenter, ask several questions of her/his interest, and to actively interact with the experimenter considering his/her interest to the addressed topics without respect to politeness. Voice data was recorded using two microphones, one headset – used in the experiments presented here – and one far-field microphone. The audio was sampled at 44.1 kHz with 16 bit. 21 subjects (ten of them female) were contained, three of them Asian, the others European. The language throughout the recordings was English as in the other two databases, and all subjects were very experienced English speakers. Three age categories were defined for balancing: younger than 30 years, younger than 40 years, and older than 40 years. The average age of male subjects was 32.7 years; the average age of female subjects was 30.1 years. The total recording time for males was 5:14:30 h, for female 5:08:00 h. In order to acquire reliable labels for the level of interest, the entire video material was segmented into speaker and sub-speaker turns and

subsequently labelled independently by four male annotators. The level of interest was annotated for every sub-speaker turn.

Five Levels of Interest were distinguished in the first place: *disinterest* (the subject is bored, listening and talking about the topic, is very passive, and does not follow the discourse), *indifference* (the subject is passive, does not give much feedback to the product presenter's explanations, and asks unmotivated questions, if any), *neutrality* (the subject follows and participates in the discourse; it cannot be determined, if she/he is interested or indifferent in the topic), *interest* (the subject wants to discuss the topic, closely follows the explanations, and asks some questions), *curiosity* (strong wish of the subject to talk and learn more about the topic). However, due to sparse occurrences, disinterest and indifference instances were combined with neutrality ones. Table 3 shows the distribution of sub-speaker-turns used for the results presented here. For details on original audiovisual class distribution, annotation and fusion of these levels of interest to a master level of interest, the reader is referred to Schuller *et al.* (2007a: 3037-3041).

Neutrality	*Interest*	*Curiosity*	*Sum*
316	510	170	996

Table 3. Distribution of speaker-turns among levels of interest in the AVIC database.

3. Acoustic features

A strictly systematic generation of features was used for the construction of a large feature space as a starting point for the subsequent selection of relevant features. Such an approach generally leads to more than 1,000 features (Vogt and Andre (2005: 474-477), Batliner *et al.* (2006: 240-245)). The basis is a set of 37 typical acoustic Low-Level-Descriptors that are known to carry information about paralinguistic effects (Batliner *et al.* (2006: 240-245), Schuller *et al.* (2007b: 2253-2256)). These descriptors are shown in Table 4. The features are split into the six commonly used types: duration

(DUR), energy (NRG), pitch (F0), formants (FX), cepstral (CEP), and voice quality (VQ). The latter is covered by jitter and shimmer (J+S), and Harmonics-to-Noise Ratio (HNR), resulting in a total of seven feature types that are investigated here. These seven types can be further grouped into three meta-groups: prosodic features, spectral features and voice quality features. The following listing provides a detailed overview regarding these three groups and explains each of the seven feature types briefly:

Prosodic features:

- *Duration (DUR):* these features model temporal aspects. Normally the basic unit is milliseconds for the 'raw' values. Thereby different types of normalisation techniques are used. The relative positions on the time axis of features like energy or pitch (maxima, on-/off-set positions, etc.) also represent duration rather than energy and pitch as such, since they are measured in milliseconds and were proven to be highly correlated with duration features in Batliner *et al.* (2001: 2781-2784). Therefore, the duration features can be distinguished regarding their extraction nature: features representing temporal properties of other acoustic base contours, and those that exclusively represent the duration of high level linguistic units such as phonemes, words, pauses, or utterances.
- *Energy (NRG):* these features model intensity, based on the amplitude, with implicit or explicit normalisation. They can model intervals or characteristic points. Since the sensation of loudness increases logarithmically as the intensity of a stimulus grows, the decibel scale is used for energy measures. The spectral distribution and the duration of a stimulus also influence loudness perception (Zwicker and Fastl, 1990). A sequence of short-term loudness values, which are extracted frame-wisely, is called a loudness contour.
- *Pitch (F0):* this is the acoustic equivalent to the perceptual unit pitch. It is measured in Hertz (Hz) and is often transformed to a perceptually more adequate representation, e.g. by mapping onto semi-tone intervals. Pitch is detected by applying a frame-based time or spectral analysis. After dividing the speech sequence into overlapping frames, the autocorrelation method, as used herein, or

spectral analysis can be applied to derive the pitch value for each frame.

Spectral features:

- *Formants (FX):* formants (i.e. spectral maxima) are known to model spoken content. This is especially true for lower order formants. Higher order formants, however, also represent speaker characteristics. Each formant is fully represented by its centre frequency, amplitude, and bandwidth. Their spectral position thereby is independent of the perceived fundamental frequency. Methods like Linear Prediction Coding (Boersma (1993: 97-110)), as used here, or cepstral analysis can be applied to estimate formant frequencies and bandwidths.
- *Cepstral (CEP):* Mel-Frequency-Cepstral-Coefficient (MFCC) features – homomorphically transformed with equidistant band-pass-filters on the Mel-frequency-scale – tend to strongly correlate with the spoken content. Yet, they have proven beneficial in practically any speech processing task. The Mel-Frequency-Cepstral-Coefficients are calculated from the logarithm of Mel-filter bank amplitudes using the discrete cosine transform.

Voice Quality features:

- *Harmonics-to-Noise Ratio (HNR):* the harmonics-to-noise ratio is a measure of the quality of the speech signal. It is estimated from voiced parts of speech and can be calculated as the ratio between the signal power in periodic parts and the signal power of the noise components. For the experiments reported here, autocorrelation was used for this calculation.
- *Jitter and Shimmer (J+S):* jitter and shimmer are micro-perturbations based on pitch and intensity reflecting voice quality.

In order to calculate Low-Level-Descriptors for the following analyses of features signalling emotion, the speech signal was transformed to 16 kHz sampling rate and 16 bit precision in the first step. In general, a Hamming window function was used, except for the calculation of pitch and Harmonics-to-Noise Ratio, where a Hanning window was

chosen. Frames were sampled at a rate of 100 Hz with 50% overlap. Thus, the effective frame length is 20 milliseconds. Energy resembles logarithmic frame energy. Pitch and Harmonics-to-Noise Ratio calculation bases on a time-signal autocorrelation with window correction. Formant calculation uses 18-point Linear Prediction Coding with root-solving and a pre-emphasis factor of 0.7. The pitch and formant trajectories were globally optimised by the use of Dynamic Programming. The Low-Level-Descriptors were smoothed by techniques such as semi-tone-interval filters or simple moving average low-pass filtering to eliminate outliers. As a next step delta (i.e. differential over time) coefficients for each Low-Level-Descriptor were added to the feature set.

Low-Level-Descriptors (2x37)	*Functionals (19)*
Pitch	Mean
Energy	Standard Deviation
Envelope	Zero-Crossing-Rate
Formant 1-5 Amplitude	Quartile 1
Formant 1-5 Bandwidth	Quartile 2
Formant 1-5 Frequency	Quartile 3
MFCC Coefficient 1-16	Quartile 1 - Minimum
Harmonics-to-Noise-Ratio	Quartile 2 - Quartile 1
Shimmer	Quartile 3 - Quartile 2
Jitter	Maximum - Quartile 3
Δ Pitch	Centroid
Δ Energy	Skewness
Δ Envelope	Kurtosis
Δ Formant 1-5 Amplitude	Maximum Value
Δ Formant 1-5 Bandwidth	Relative Maximum Position
Δ Formant 1-5 Frequency	Minimum Value
Δ MFCC Coefficient 1-16	Relative Minimum Position
Δ Harmonics-to-Noise-Ratio	Maximum Minimum Range
Δ Shimmer	Position of 95% Roll-Off-Point
Δ Jitter	

Table 4. Low-Level-Descriptors and Functionals considered for emotion pair discrimination.

Next, the typical static classification strategy commonly used in emotion recognition, e.g. by Batliner *et al.* (2006: 240-245), was employed: a total of 19 statistical functionals was applied to each of the 2x37 Low-Level-Descriptors. Thus, the obtained multivariate time

series of variable length was projected onto a single 1,406 dimensional feature vector. A typical selection of common functionals covering the first four statistical moments, quartiles, extremes, ranges, positions, and zero-crossings, was chosen, as depicted in Table 4. The three position related functionals lead to a sub-group of features with the physical unit of milliseconds which were treated as duration features, though being based on a number of diverse Low-Level-Descriptors as described above. Further, duration related features such as those based on e.g. lengths of pauses or syllables were not included because this information cannot be easily integrated in the strictly systematic generation approach (they are modelled in a general value series rather than in a time series).

Table 5 shows the obtained distribution of features among the introduced types. As can be seen, considerably more spectral than prosodic features were extracted. The fewest features were extracted for voice quality. However, this reflects the standard situation (Batliner *et al.* (2006: 240-245)): cepstral coefficients and formants result in a higher number of Low-Level-Descriptors as compared to pitch and energy, for example.

Group	*DUR*	*NRG*	*F0*	*FX*	*CEP*	*HNR*	*J+S*
Features [#]	222	64	32	480	512	32	64
Relative [%]	15.8	4.6	2.3	34.1	36.4	2.3	4.6
Type	*Prosodic*			*Spectral*		*Voice Quality*	
Features [#]	318			992		96	
Relative [%]	22.6			70.6		6.8	

Table 5. Distribution of frequency of acoustic features among types. DUR: duration, NRG: energy, F0: pitch, FX: formants, CEP: cepstral, HNR: Harmonics-to-Noise Ratio, J+S: jitter and shimmer.

The next section deals with the technique used to automatically find relevant features to distinguish affect/emotion pairs. The results, i.e. the features found to be most relevant, are discussed in Section 5.

4. Classification and feature selection

In order to automatically recognise affect using acoustic features, advanced classification algorithms are required. The classification method used in this chapter is called Support-Vector-Machines (SVM). The model for the Support-Vector-Machines is built using a Sequential Minimal Optimisation (SMO) Algorithm (Witten, 2005). The Support-Vector-Machine classifier is a binary classifier, i.e. only two classes can be separated with one model. However, as can be seen in Section 2, usually more than two classes are required for recognition of emotion/affect. For multi-class problems a technique called Round Robin, i.e. one-against-one, classification is used (Fürnkranz (2002: 721-747)). If n is the total number of classes there are $n(n-1)/2$ class pairs. For each pair a model is built that is able to distinguish only between these two classes. This classification technique was used for the work described herein.

Classification and modelling techniques are only the second part of the two essential parts of machine learning. The first part is the selection of appropriate features. "Appropriate" in this context refers to being highly predictive, i.e. attributes that are highly correlated with the classes that are to be separated. For emotion recognition based on large sets of acoustic features the attribute selection, besides improving classifier performance, also serves the purpose of reducing the number of required features, which results in faster processing. The selected attributes are furthermore of interest in paralinguistic and prosodic studies since they are clues as to how affect and emotion is encoded in spoken language. Given a set of more than a thousand statistical features computed from the acoustic Low-Level-Descriptors as introduced in Section 3, feature selection is not straight forward; preferably, automatic feature selection methods are used (Schuller *et al.* (2005: 805-809)).

Generally, a target function is required as optimisation criterion for reduction of the feature space, meaning the reduction of the dimensionality of the feature vector (the number of features). An attribute evaluator based on Correlation-based Feature Subset Selection was used (Witten, 2005).

Correlation-based Feature Subset Selection evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred. Since the feature set is very large, not all possible combinations of feature subsets can be evaluated in a reasonable amount of time. Therefore – as also stated in other works on speech emotion recognition (Ververidis and Kotropoulos (2005: 1500-1503), (2006: 1162-1181)) – an exhaustive search is not an efficient and applicable option. Instead, a Sequential Forward Floating Search (SFFS) (Pudil *et al*. (1994: 1119-1125)) – the most commonly used search function in this field (Lee and Narayanan (2005: 293–303), Ververidis and Kotropoulos (2005: 1500-1503), Vogt and Andre (2005: 474-477)), Schuller *et al.* (2005: 805-809)) which is also highly competitive in general feature selection (Zongker (1997: 18-22)) – was applied. This approach refers to the space of feature subsets being processed by greedy hill climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allows for control of the level of backtracking being done. Generally, Sequential Forward Floating Search may start with the empty set of features and search forward (known as best first), start with the full set of features and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point). Here, the forward search direction was chosen since a small number of final features were expected in order to keep the complexity low and increase the processing speed.

Since Correlation-based Feature Subset Selection is an automatic and unsupervised feature selection method based on statistical properties of the data, the resulting feature sets are potentially not the general best set for every possible classifier or the overall most representative sets of attributes that encode affect in speech communication. To evaluate which feature types are essential for detecting specific affective states or emotions, binary classification test-runs were conducted for every possible class pair in each database using only one of the seven feature type sets described in Section 3. These test-runs were performed using a) the full (not reduced) set of 1,406 features and b) a set reduced by Correlation-based Feature

Subset Selection. Both sets were split into subsets containing one of the seven feature types.

The evaluations of the test-runs were performed in a Leave-One-Speaker-Out speaker independent cross validation. 'Speaker independent' implies that the speaker of the classified utterances is not included in the training data set. She is unknown to the classifier and the deduced learning rules (Lugger and Yang (2007: 2149-2152)), which is important for realistic use-cases (Schuller *et al.* (2007c: 941-944)). The results are reported in form of mean accuracy (RR), non-weighted mean recall (CL) and the harmonic mean F of mean accuracy and mean recall as given by equation 1.

$$F = \frac{2RR \cdot CL}{RR + CL}$$

Equation 1. Harmonic mean F.

These three different measures are necessary in case the number of instances per class is not distributed equally among the classes, as is usually the case for emotional speech corpora (see Section 2). Results for classes with fewer instances tend to be below the actual achievable result for that class. This is due to insufficient amount of training data and low a-priori probability for the class. Likewise, the classes with more instances have a higher a-priori probability and a more robustly estimated model due to the higher amount of available training data. The recognition performance for such a class is usually over-estimated. Classes with many instances have a substantial influence on the RR measure, whereas smaller classes have very little influence. Thus, the RR measure tends to over-estimate performance when the instance numbers of the classes are very unbalanced. In the same way, not weighting the mean recall rate with the number of instances in each class increases the influence of classes with very few instances, thus under-estimating overall performance. Therefore, the F-measure is commonly used to give an estimate of a balanced result even if the size of the classes is not balanced (Batliner *et al.* (2006: 240-245), Schuller *et al.* (2007c: 941-944)).

5. Feature type relevance

This section presents the results for the experiments described in the previous section. As mentioned in the last section, Leave-One-Speaker-Out evaluation is used for subject independent testing. Mean results over all subjects are provided. In Table 6 the relevance of the non-reduced feature types (i.e. using all 1,406 features split into the seven types) for all possible emotion pairs can be seen. In addition to the performance measures RR and CL, the F-measure is given as described in Section 4. Further, emotion pairs are listed in Table 6 for the three cases 'Big Six', stress, and interest levels, covered by the databases eNTERFACE, SUSAS, and AVIC. The part on the right of the Table 6 shows the absolute difference between the mean F-measure and the F-measure for each Low-Level-Descriptor type, which is referred to as ΔF here.

For 11 out of 24 possible emotion pairs, the best performance can be obtained when using formant features (FX). Cepstral features (CEP) outperformed all other Low-Level-Descriptor types in 12 cases. Only in one case (emotion pair fear/surprise, eNTERFACE database) did the usage of pitch features (F0) yield the best results. Considering the emotion pairs of the eNTERFACE database, the performance gain with respect to the mean F-value varied between 1.7% and 12.4% when using formant features as Low-Level-Descriptors, and between 0.9% and 12.8% when using cepstral features. When classifying the emotion pairs of the SUSAS database, cepstral features outperformed formant Low-Level-Descriptors for all emotion pairs. Contrariwise formant Low-Level-Descriptors prevailed for interest-related affect in the AVIC database.

Consequently, in 23 out of 24 cases the best emotion recognition performance could be obtained with spectral features alone. Averaged over all emotion pairs, the F-measure for all prosodic and voice quality Low-Level-Descriptors was lower than the mean F-measure (negative mean ΔF), while the mean ΔF was 6.9% for formant features and 7.1% for cepstral features respectively.

[%]	*RR*	*CL*	*F*	*Δ F*						
	Mean			*DUR*	*NRG*	*F0*	*FX*	*CEP*	*HNR*	*J+S*
anger/disgust	73.4	71.2	72.1	0.8	-3.0	-5.5	**8.1**	7.6	-3.0	-5.0
anger/fear	72.4	71.7	72.1	0.1	2.3	-3.0	**8.0**	6.8	-6.0	-8.2
anger/happiness	75.1	71.0	74.8	-3.3	0.6	-2.2	**7.5**	5.3	-3.7	-4.2
anger/sadness	76.4	69.3	76.0	-6.6	-2.1	0.9	10.5	**12.8**	-7.7	-7.7
anger/surprise	74.8	73.2	73.8	0.5	2.5	-6.9	8.9	**9.7**	-9.7	-5.0
disgust/fear	66.4	63.2	66.0	-2.4	-2.9	6.4	**6.8**	0.9	-0.6	-8.2
disgust/happiness	62.2	59.7	61.4	-1.0	1.4	-1.1	3.1	**3.5**	-2.1	-3.8
disgust/sadness	69.4	61.1	68.3	-5.8	-1.6	4.3	5.6	**12.2**	-7.3	-7.4
disgust/surprise	65.6	62.2	64.7	-0.6	-7.1	-0.4	6.0	**10.6**	-4.3	-4.3
fear/happiness	69.3	67.1	69.3	-2.7	-0.3	1.4	**7.0**	6.0	-3.1	-8.5
fear/sadness	66.0	65.1	65.0	1.0	-1.7	2.0	**10.3**	4.6	-5.1	-11.1
fear/surprise	63.7	64.2	63.4	1.4	-8.0	**6.3**	1.7	5.1	-1.7	-4.9
happiness/sadness	72.5	67.6	71.6	-3.6	0.7	3.5	**12.4**	8.1	-12.0	-9.1
happiness/surprise	65.8	64.4	65.5	-0.7	-2.1	-1.7	6.3	**8.2**	-7.8	-2.3
sadness/surprise	70.1	69.5	68.8	2.3	-5.1	5.1	**9.5**	4.9	-9.2	-7.4
high stress/med.	58.3	57.0	58.3	-1.4	2.0	0.1	-0.7	**3.5**	-4.0	0.7
high stress/neutral	59.7	53.4	56.5	-0.7	-1.0	-4.0	-2.5	**8.3**	-0.6	0.1
high stress/fear	89.3	80.1	86.0	-3.1	5.5	-4.5	11.1	**11.4**	-18.3	-3.0
med.stress/neutral	61.5	56.0	56.9	2.3	1.0	-0.4	-0.3	**4.5**	-8.1	-0.2
med.stress/fear	90.6	82.9	87.4	-1.9	6.0	-4.9	10.7	**11.1**	-14.1	-7.5
neutral/fear	87.8	82.3	87.0	-3.8	4.7	-2.5	10.7	**11.5**	-13.8	-6.9
neutral/interest	74.5	75.2	73.4	1.9	3.1	-1.2	**2.4**	1.2	-3.0	-3.5
neutral/curiosity	75.2	73.2	74.1	-0.2	1.5	-4.2	**14.6**	10.7	-13.3	-9.4
interest/curiosity	73.9	60.0	64.5	0.9	-1.4	-2.9	**8.4**	2.5	-4.5	-4.7
mean	71.4	67.5	69.9	-1.1	-0.2	-0.6	6.9	**7.1**	-6.8	-5.5

Table 6. Relevance of non-reduced Low-Level-Descriptor types for all emotion pairs by mean accuracy (RR), non-weighted recall (CL) and harmonic mean of these two (F). Further the absolute difference for each Low-Level-Descriptor type to the mean is provided (ΔF). Databases in order of appearance: eNTERFACE (top), SUSAS (middle), AVIC (bottom). Speaker-independent Leave-One-Speaker-Out evaluation. Support-Vector Machine classification.

Using only voice quality features (HNR and J+S) led to the worst performance (ΔF of -6.8% and -5.5% respectively).

For the emotions in the eNTERFACE database no relation between the emotion pair and whether cepstral or formant features are

more relevant could be found. However, an interesting fact can be observed when looking at the ΔF measure of pitch. If sadness or fear is one part of the emotion pair (except for anger/fear), the pitch features performed better than the mean (i.e. ΔF is greater than zero), yet still worse than the spectral features. The latter can be explained because information about pitch can also be interpreted as spectral information. It can be assumed that the spectral features to some extent contain the same information as carried by the pitch features alone, but add more information about higher level harmonics. The fact that for fear and especially sadness pitch performed better than the average may indicate that pitch variations make a person sound uneasy. When a person is described as speaking fearfully or sadly, the term 'trembling voice' is often used to refer to pitch modulations.

[%]	*DUR*	*NRG*	*F0*	*FX*	*CEP*	*HNR*	*J+S*	*Pros*	*Spec*	*VQ*
'Big Six'	13.6	3.4	3.4	33.0	45.5	0.0	1.1	20.4	78.5	1.1
Stress	1.2	7.0	5.9	31.8	48.2	0.0	5.9	14.1	80.0	5.9
Interest	4.1	7.3	1.6	37.4	49.6	0.0	0.0	13.0	87.0	0.0

Table 7. Relevance of Low-Level-Descriptor types and groups (Pros: prosodic, Spec: spectral, VQ: voice quality) averaged over all emotion pairs by percentage within the final selected subset after Correlation-based Feature Subset Selection with Sequential Forward Floating Search. Databases in order of appearance: eNTERFACE (top), SUSAS (middle), AVIC (bottom).

The importance of spectral features for affect recognition can also be seen in Table 7, which shows the results of feature selection. Averaged over all emotion pairs, the percentage within the final feature subset after Sequential Forward Floating Search is given for every database as well as for every Low-Level-Descriptor type (on the left) and Low-Level-Descriptor group (on the right). In the case where feature selection was carried out for every Low-Level-Descriptor type, between 45.5% and 49.6% of the remaining features were cepstral Low Level Descriptors. The percentage of remaining formant features varied between 33% and 37.4%, depending on the database. Only a small fraction of the selected features were prosodic or voice quality Low Level Descriptors. The Harmonics-to-Noise Ratio features were

found not to be relevant at all for signalling affect, regardless of the database.

[%]	*DUR*	*NRG*	*F0*	*FX*	*CEP*	*HNR*	*J+S*	*Pros*	*Spec*	*VQ*	*All*
'Big Six'											
without Feature Selection											
RR	68.3	67.6	70.2	77.2	76.6	63.9	62.9	72.8	80.3	65.0	81.9
CL	66.7	66.6	68.7	75.5	75.3	62.7	61.9	70.0	79.0	62.7	80.6
F	67.5	67.1	69.4	76.3	75.9	63.3	62.4	71.4	79.6	63.8	81.2
with Feature Selection											
RR	69.9	68.3	69.3	75.5	76.6	63.1	62.9	76.5	81.4	63.9	83.1
CL	68.7	67.1	68.0	74.1	75.7	61.8	61.9	75.1	80.2	62.8	81.8
F	69.3	67.7	68.6	74.8	76.1	62.4	62.4	75.8	80.8	63.3	82.4
Stress											
without Feature Selection											
RR	72.8	76.8	72.6	77.1	80.8	67.8	73.7	77.1	79.3	74.5	80.2
CL	68.6	73.5	66.5	76.6	80.0	57.7	65.5	75.2	79.0	68.4	80.2
F	70.6	75.1	69.4	76.8	80.4	62.3	69.4	76.1	79.1	71.3	80.2
with Feature Selection											
RR	73.7	77.8	73.5	77.3	82.1	67.4	74.4	78.2	80.5	73.1	80.5
CL	67.2	72.3	66.6	75.6	80.2	55.8	61.4	74.9	79.1	65.1	79.0
F	70.3	74.9	69.9	76.4	81.1	61.1	67.3	76.5	79.8	68.9	79.7
Interest											
without Feature Selection											
RR	74.0	76.1	73.5	80.1	77.0	70.8	71.1	78.2	79.2	70.6	69.5
CL	69.5	68.3	63.7	78.3	74.1	58.7	60.3	75.8	76.5	60.3	69.3
F	71.7	72.0	68.3	79.2	75.5	64.2	65.3	77.0	77.8	65.0	69.4
with Feature Selection											
RR	76.2	76.7	71.0	83.3	81.6	68.8	69.0	78.3	84.2	69.9	74.1
CL	66.4	68.2	58.1	80.1	77.9	56.1	55.9	70.1	81.8	57.1	75.8
F	71.0	72.2	63.9	81.7	79.7	61.8	61.8	74.0	83.0	62.9	74.9

Table 8. Relevance of Low-Level-Descriptor types and groups and accuracies of all features (Pros: prosodic, Spec: spectral, VQ: voice quality) averaged over all emotion pairs by mean accuracy (RR), non-weighted recall (CL) and harmonic mean of these two (F) with and without Feature Selection. Databases in order of appearance: eNTERFACE (top), SUSAS (middle), AVIC (bottom).

Feature selection was also carried out for each Low-Level-Descriptor type. Again, the majority of the selected features were spectral

attributes (between 78.4% and 87%). Compared to the recognition of the 'Big Six' emotions of the eNTERFACE database, prosodic features seem to be slightly more relevant for interest-related user states as in the AVIC database. Here, 20.5% of the remaining features were prosodic Low-Level-Descriptors. In this respect the initially different amount of features per group to pick from has to be considered.

Table 8 shows the recognition results for the three emotion sets with and without feature selection. As before, best results could be obtained when using spectral features. Thereby, feature selection carried out with the spectral Low Level Descriptors, slightly improved recognition performance. Using only spectral features resulted in a performance almost as good as obtained when incorporating all features. For the interest-related emotions the usage of spectral features alone even outperformed classification with all features. This demonstrates that non-exhaustive feature selections simply are sub-optimal and can lead to lower performance in rare cases.

6. Conclusion

Aiming to determine features in the speech signal which are optimally suited to represent affect in human speech, several experiments were performed to automatically discriminate between varieties of emotion pairs. In total, 1,406 features were extracted from the audio data taken from three databases (eNTERFACE, SUSAS, AVIC). The experiments proved that spectral features contain the most relevant information about emotion within speech for almost every pair of emotions and independent of the dataset. When using only formant features for emotion recognition, performance was shown to be 6.9% higher compared to the average. Cepstral features performed even 7.1% better than the average feature type. Pitch, as a prosodic Low-Level-Descriptor, prevailed only for the emotion pair fear/surprise where it performed 6.3% better than the average. Considering the different emotion sets, cepstral features were proved to be best suited

for the discrimination of stress related emotions whereas formants outperformed all other feature types for emotions expressing interest. Emotion pairs like high stress/medium stress, medium stress/neutral, or fear/surprise were more difficult to classify than pairs such as neutral/fear or anger/sadness.

The importance of spectral features cannot only be seen in recognition performance but also in the outcome of Correlation-based Feature Subset Selection: depending on the database, between 78.4% and 87% of the remaining features were spectral attributes whereas voice quality features were almost completely rejected. Thereby it has to be considered that spectral features also encode linguistic information. This is not of importance in the case of the eNTERFACE and SUSAS sets, where the spoken content is predefined and fixed for different emotions. However, in the case of the AVIC set this may well play a further important factor.

The feature selection could slightly improve recognition results when using only spectral features. Thereby the usage of spectral features performed almost as well as a classifier operating with all feature types – even better for interest related emotions.

All conducted emotion recognition experiments highlight the high importance of spectral features such as formants and cepstral coefficients and prove that most information about emotion within speech can be found directly in the spectrum of the speech signal and in prosodic features – as opposed to many earlier works in which the first type plays only a secondary role (e.g. Ang *et al.* (2002: 2037-2040)). At the same time voice quality-based features were not found to be of high importance, as opposed to e.g. Scherer (1986: 143–165). However, works in favour of voice quality features often derive their findings from listening experiments with acted or synthesised data (c.f. Gobl *et al.* (2003: 189–212)). Speaker-independent automatic recognition of affect clearly differs from these settings. Finally, prosodic features clearly helped to improve overall performance, though on their own they were not the best choice. Similar to the voice quality features, their assumed importance often stems from emotional voice synthesis and listening experiments. Thereby these parameters are easier to edit, better understood and intuitive for listeners than the more abstract and complex spectral characteristics.

While future research efforts will have to verify these findings in further corpora, the discussed findings clearly support the trends towards expecting that among other current changes in the field of emotional speech analysis (e.g. Wöllmer *et al.* (2008: 597–600), Schuller *et al.* (2007c: 1500-1503)) a shift in the preferred feature type will be observable.

References

i) books

Witten, Ian H. / Frank, Eibe. 2005: *Data Mining: Practical Machine Learning Tools with Java Implementations*, 2nd edition. San Francisco: Morgan Kaufmann.

Zwicker, Eberhard. / Fastl, Hugo 1990. *Psychoacoustics. Facts and Models.* Series in Information Sciences, Volume 22. Berlin: Springer-Verlag.

ii) articles in books:

Ang, Jeremy / Dhillon, Rajdip / Krupski, Ashley / Shriberg, Elizabeth / Stolcke, Andreas 2002. Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog, In *Proceedings of ICSLP 2002*. Denver, Colorado. 2037-2040.

Batliner, Anton / Steidl, Stefan / Schuller, Björn / Seppi, Dino / Laskowski, Kornel / Vogt, Thurid / Devillers, Laurence / Vidrascu, Laurence / Amir, Noam / Kessous, Loic / Aharonson, Vered 2006. Combining Efforts for Improving Automatic Classification of Emotional User States. In *Proceedings of IS-LTC 2006*. Ljubliana, 240-245.

Batliner, Anton / Schuller, Björn / Schaeffler, Sonja / Steidl, Stefan 2008. Mothers, Adults, Children, Pets - Towards the Acoustics of Intimacy. In *Proeedings of Int. Conf. on Acoustics, Speech, and*

Signal Processing ICASSP 2008, Las Vegas, Nevada. Washington: IEEE, 4497-4500.

Boersma, Paul 1993. Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound. In *Proceedings of the Institute of Phonetic Sciences*, Volume 17. Amsterdam. 97-110.

Ekman, Paul 1972. Universals and Cultural Differences in Facial Expressions of Emotion. In Cole, J. (eds.) *Nebraska Symposium on Motivation 1971, Volume 19*. Lincoln: University of Nebraska Press, 207-283.

Hansen, John H. L. / Bou-Ghazale, Sahar 1997. Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In *Proceedings of EUROSPEECH-97*, *Volume 4. Rhodes, Greece*. Lisbon: ISCA, 1743-1746.

Lugger, Marco / Yang, Bin 2007. An Incremental Analysis of Different Feature Groups in Speaker Independent Emotion Recognition. In *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbruecken, Germany*. Washington: IEEE, 2149-2152.

Martin, Olivier / Kotsia, Ioannis / Macq, Benoît / Pitas, Ioannis 2006. The Enterface'05 Audio-Visual Emotion Database. In *Proceedings of the IEEE Workshop on Multimedia Database Management, Atlanta*. Washington: IEEE. 207-212.

McGilloway, Sinéad / Cowie, Roddy / Douglas-Cowie, Ellen / Gielen, Stan / Westerdijk, Machiel / Stroeve, Sybert 2000. Approaching Automatic Recognition of Emotion from Voice: a Rough Benchmark. In *Proceedings of the ISCA Workshop on Speech and Emotion.* Lisbon: ISCA, 207-212.

Schuller, Björn / Müller, Ronald / Lang, Manfred / Rigoll, Gerhard 2005. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. In *Proceedings of the 9th Eurospeech - 6th Interspeech 2005.* Lisbon: ISCA, 805-809.

Schuller, Björn / Müller, Ronald / Hörnler, Benedikt / Höthker, Anja / Konosu, Hitoshi / Rigoll, Gerhard 2007a: Audiovisual recognition of Spontaneous Interest within Conversations. In *Proceedings of the 9th Int. Conference on Multimodal Interfaces (ICMI)*, Special Session on Multimodal Analysis of Human Spontaneous Behaviour. Nagoya, Japan: ACM SIGCHI. 3037-3041.

Schuller, Björn / Batliner, Anton / Seppi, Dino / Steidl, Stefan / Vogt, Thurid / Wagner, Johannes / Devillers, Laurence / Vidrascu, Laurence / Amir, Noam / Kessous, Loic / Aharonson, Vered 2007b. The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. In *Proceedings 8th Interspeech 2007, Antwerp*. Lisbon: ISCA, 2253-2256.

Schuller, Björn / Seppi, Dino / Batliner, Anton / Maier, Andreas / Steidl, Stefan 2007c. Towards More Reality in the Recognition of Emotional Speech. In *Proceedings of ICASSP 2007*, *Honolulu, Hawaii*. Washington: IEEE. 941-944.

Schuller, Björn / Rigoll, Gerhard / Can, Salman / Feussner, Hubertus 2008. Emotion Sensitive Speech Control for Human-Robot Interaction in Minimal Invasive Surgery. In *Proceedings 17th Intern. Symposium on Robot and Human Interactive Communication RO-MAN 2008, Munich.* Washington: IEEE, 453–458.

Ververidis, Dimitrios / Kotropoulos, Constantine 2005. Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm. In *Proceedings of ICME 2005, Amsterdam*. Washington: IEEE, 1500-1503.

Vlasenko, Bogdan / Schuller, Björn / Wendemuth, Andreas / Rigoll, Gerhard 2007. Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. In *Proceedings 2nd Int. Conf. on Affective Computing and Intelligent Interaction ACII 2007, Lisbon, Portugal*, volume LNCS 4738. Berlin/Heidelberg: Springer. 139-147.

Vogt, Thurid / Andre, Elisabeth 2005. Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. In *Proceedings of ICME 2005*, Amsterdam. Washington: IEEE, 474–477.

Wöllmer, Martin / Eyben, Florian / Reiter, Stephan / Schuller, Björn / Cox, Cate / Douglas-Cowie, Ellen / Cowie, Roddy 2008. Abandoning Emotion Classes – Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies. In *Proceedings 9th Interspeech 2008, Brisbane*. Lisbon: ISCA. 597–600

Zongker, Douglas / Jain, Anil 1996. Algorithms for Feature Selection: An Evaluation. In *Proceedings of the International Conference on Pattern Recognition* (ICPR 96)*, Vienna, Austria*. Washington: IEEE, 18-22.

iii) articles in journals:

Cowie, Roddy / Douglas-Cowie, Ellen / Tsapatsoulis, Nicolas / Votsis, George / Kollias, Stefanos / Fellenz, Winfried / Taylor, John G. 2001. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing magazine* 18/1. 32-80.

Fürnkranz, Johannes 2002. Round Robin Classification. *Journal of Machine Learning Research* 2. 721-747.

Gobl, Christer / Ní Chasaide, Ailbhe 2003. The Role of Voice Quality in Communicating Emotion, Mood and Attitude. *Speech Communication* 40, 189-212.

Lee, Chul Min / Narayanan, Shrikanth 2005. Toward Detecting Emotions in Spoken Dialogs. *IEEE Transactions on Speech and Audio Processing* 13(2), 293-303.

Pantic, Maja / Rothkrantz, Leon J. M. 2003. Toward an Affect-Sensitive Multimodal Human-Computer Interaction, *Proceedings of the IEEE* 91, 1370-1390.

Scherer, Klaus R. 1986. Vocal Affect Expression: a Review and a Model for Future Research. *Psychological Bulletin* 99, 143-165.

Ververidis, Dimitrios / Kotropoulos, Constantine 2006. Emotional Speech Recognition: Resources, Features, and Methods. *Speech Communication* 48/9*,* 1162-1181.