CHAPTER 16

Emotion Estimation in Speech Using a 3D Emotion Space Concept

Michael Grimm, Kristian Kroschel
Universität Karlsruhe (TH)
Germany

1. Introduction

Automated recognition of emotions conveyed in the speech is an important research topic that has emerged in recent years with a number of possible applications (Picard, 1997). The most important one is probably the improvement of the man-machine interface, knowing that human communication contains a large amount of emotional messages which should be recognized by machines such as robot assistants in the household, computer tutors, or automatic speech recognition (ASR) units in call-centers.

Most research on recognizing emotions in the speech focuses on a small number of emotion categories (Dellaert et al., 1996; Lee et al., 2001; Yu et al., 2004; Vidrascu & Devillers, 2005; Schuller et al., 2005). However, such categorization is a strong restriction if we think of the continuum in the expression of human emotions. In particular, if we want to resolve moderate emotions in spontaneous utterances in addition to the stereotype portrayal of exaggerated emotions, several questions must be asked:

- How can we estimate the emotion conveyed in the speech signal more detailed than in the state-of-the-art categorization?
- How many features are necessary for such estimation, and which method is suitable for selection?
- Which estimation methods are suitable, and what are the pros and the cons of each method?

In this contribution we intend to give some answers to these questions, building upon our previous work reported in (Grimm et al., 2007a; 2007b). We propose a generalized framework using a continuous-valued, three-dimensional emotion space method. This method defines emotions as points in a three-dimensional emotion space spanned by the three basic attributes ("primitives") *valence* (positive-negative axis), *activation* (calm-excited axis), and *dominance* (weak-strong axis) (see Kehrein, 2002). Figure 1 shows a schematic sketch of this emotion space. *Anger,* e.g., would be represented by negative *valence,* high excitation level on the *activation* axis, and strong *dominance*. Such real-valued notion was shown to be transferable to emotion categories, if desired (Grimm et al., 2006). At the same time, it lends itself to a gradual description of emotion which is helpful to describe intensity changes over time or speaker-dependent emotion expression behaviors, for example.

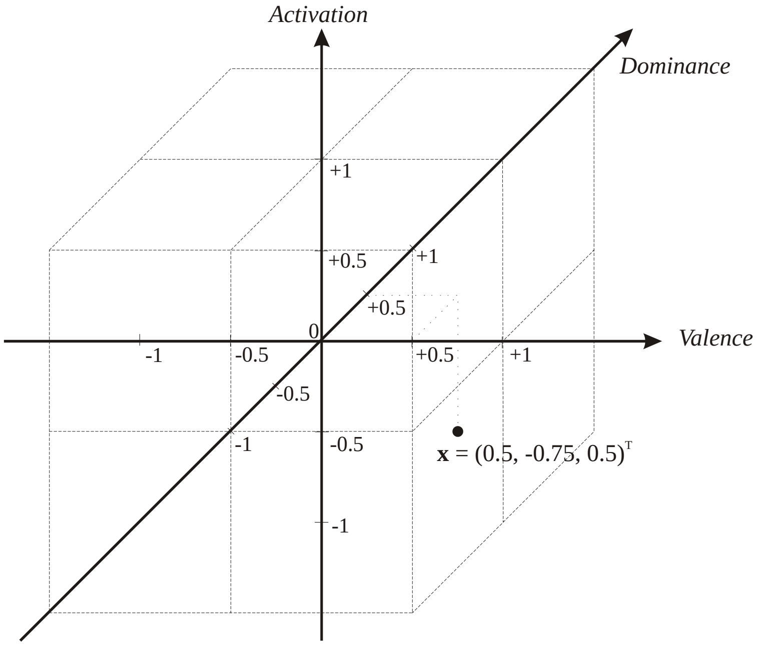

Figure 1: Three-dimensional emotion space, spanned by the primitives *valence*, *activation*, and *dominance,* with a sample emotion vector added for illustration of the component concept.

While emotion space concepts have been named as an alternative to emotion categories in many studies (see Cowie, et al., 2001 or Cowie & Cornelius, 2003 for a comprehensive overview), they have so far only been used for emotional speech synthesis (Schröder, 2003), and hardly for emotion recognition. In the few other studies on detecting emotions in speech using an emotion space concept, first the emotion space is subdivided into 2 or 3 subregions and then these emotion space regions are classified just as emotion categories (Fragopanagos & Taylor, 2005; Vidrascu & Devillers, 2005). Thus, we consider the direct, continuous-valued estimation of these emotion primitives in this contribution.

Based on a German speech database containing authentic emotional expressions from a TV talk-show, the automatic estimation of these three emotion primitives was studied. The reference was built by a human listener test using the text-free method of Self Assessment Manikins (see Section 2.2). The speech signal was segmented into utterances for this study yielding 893 samples of 47 speakers. For the automatic estimation of the emotion in these utterances, a feature vector was built containing the statistics of the fundamental frequency, energy and the MFCCs, such as mean value, standard deviation, minimum, maximum, range, and quartiles. Two different techniques to reduce the feature vector size were studied. As an automatic emotion estimation method we used Support Vector Regression (SVR), which provides a kernel-based, non-linear estimation of the three emotion primitives. This method showed promising results and a moderate to high correlation with the human listeners' ratings in a preliminary study. It is now compared to KNN and Fuzzy Logic both with respect to the estimation error and the learning curve properties (see Sections 5.1 and 5.2, respectively).

The rest of the paper is organized as follows. Section 2 briefly introduces the data we used, and it also describes the emotion evaluation by human listeners. Section 3 describes the pre-processing steps of feature extraction and feature selection. Section 4 presents the different classifiers used for continuous-valued emotion primitive estimation. Section 5 describes the results and discusses the different estimator outcomes. Section 6 contains the conclusion and directions for future work.

The term "emotion" is very difficult to define. It refers to a very complex inner state of a person including a wide range of cognitive and physical events (Scherer, 1990). In addition to the truly felt affective state, the expression of emotions is superimposed by the display rules of the situation. In the scope of this chapter we understand the term "emotion" only as the visible part of this inner state that is transmitted through the speech signal and thus observable by a human receiver. Also, the conclusions drawn from the automatic estimation have to be seen in the context of the situation. Analyzing additional channels, such as the mimics, gestures, or physiological signals, might improve the understanding of the situation and thus help identify the emotional state of a person more precisely.

2. Data

2.1 Data acquisition

To study the emotion recognition on authentic emotions in speech we extracted dialogue episodes from a talk-show on TV. In this talk-show, two or three persons discuss problems such as fatherhood questions, friendship issues, or difficulties in the family. Due to the spontaneous and unscripted manner of the episodes, the emotional expressions can be considered authentic. Due to the topics, the data contains many negative emotions and few positive ones.

This data was first introduced as *VAM corpus*[1] in (Grimm & Kroschel, 2005b). All signals were sampled at 16 kHz and 16 bit resolution. In total the corpus contains 893 sentences from 47 speakers (11m/36f).

The emotion in each utterance was evaluated in a listener test (c.f. Section 2.2). Based on such a human evaluation, Figure 2 shows the histogram of the emotions contained in the database. The attested emotion was taken as the reference for the automatic recognition, since assessment by the speakers themselves was not available.

[1] The acronym VAM resides from the title of the talk-show, "Vera am Mittag" (German: Vera at Noon).

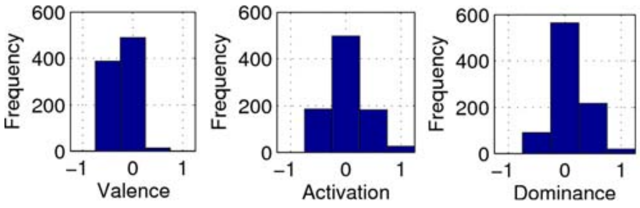

Figure 2: Distribution of the emotions present in the VAM corpus.

2.2. Emotion evaluation

For evaluation we used a listener test. A group of evaluators listened to the emotional sentences and assessed the emotional content. For this human evaluation we used the Self Assessment Manikins (Fischer, et al., 2002; Grimm & Kroschel, 2005a). In this method five images were offered per emotion primitive. For each sentence, an evaluator listened to the speech signal. Then he/she was asked to select the best describing image for each primitive. This evaluation method yields one reference value $x_n^{(i)} \in \{-1, -0.5, 0, 0.5, 1\}$ for each primitive $i \in$ {*valence, activation, dominance*} and each utterance n. The individual listener ratings were averaged using confidence scores as described in (Grimm & Kroschel, 2005a).

One half of the database was evaluated by 17 listeners, the other by 6 listeners, which was due to the fact that the second half of the database was recorded and evaluated later, when only a smaller number of evaluators was available. In comparison with other studies on emotion recognition which included 2 to 5 independent evaluations (Vidrascu & Devillers, 2005; Yu et al., 2004), we used a much higher number of evaluators to gain statistical confidence.

The average standard deviation in the evaluation was 0.29, 0.34, and 0.31 for *valence*, *activation*, and *dominance*, respectively. Thus, the average human deviation in the evaluation test was slightly above one half of the distance between two images, which was notably low for such a difficult task. The mean correlation between the evaluators was 0.49, 0.72, and 0.61, respectively (Grimm et al., 2007a), measured by Pearson's empirical correlation coefficient. Thus, *valence* was significantly more difficult to evaluate than *activation* or *dominance*. However, this result might also be an artifact of the correlation coefficient including the variance of the distribution, which is also smaller for *valence*.

3. Pre-processing

3.1. Feature extraction

The speaker's emotion is conveyed through a number of different channels. The most apparent correlates are found in the prosody of the speech. An overview on the acoustic expression of emotions can be found in (Murray & Arnott, 1993; Cowie & Cornelius, 2003). Analyzing the linguistic content provided by ASR in addition to the prosody might improve the estimation of the emotion (Lee & Narayanan, 2005). However, it has to be kept in mind that the performance of the ASR degrades remarkably in the case of emotional speech. Therefore, we concentrate on the non-linguistic information in the speech signal in this study.

In accordance with other research on automatic emotion recognition, we extracted prosodic features from the fundamental frequency (pitch) and the energy contours of the speech signals. The first and the second derivatives were also used. For pitch extraction, the autocorrelation method was used. For each of these 6 signals (pitch, energy x 0th, 1st, 2nd derivative) we calculated the following 9 statistical parameters:

- mean value
- standard deviation
- median
- minimum (not for energy)
- maximum
- 25% quantile
- 75% quantile
- difference between maximum and minimum
- difference between the quartiles

In addition we used 6 temporal characteristics:

- pause-to-speech ratio
- speech duration mean
- speech duration standard deviation
- pause duration mean
- pause duration standard deviation
- speaking rate

Finally, the spectral characteristics were added to the feature set. The Mel Frequency Cepstral Coefficients (MFCCs) were calculated in 13 subbands. The increased bandwidth with increasing center frequency of the individual subbands thereby reflects the hearing characteristics of the human ear (O'Shaughnessy, 1999). The mean value and the standard deviation of each of the 39 MFCC trajectories (13 subbands x 0th, 1st, 2nd derivative) were added to the feature set.

Thus, in total 137 acoustic features were extracted: 53 features derived from pitch and energy, 6 temporal characteristics, and 78 features to describe the spectral variability in the cause of the utterance. They were normalized to the range [0, 1].

3.2. Feature selection

To reduce the large amount of acoustic features, we used two different methods: Sequential Forward Selection (SFS) and Principal Component Analysis (PCA).

In the Sequential Forward Selection (SFS) technique for feature selection (Kittler, 1978), the feature set is increased sequentially. In each iteration, the one feature is added to the set which minimizes the classification error. Listing 1 contains the pseudo-code for this procedure.

```
current_best_feature_set := { }
remaining_features := {feat1, ..., feat137}
for num_features = 1 to 137
  error := { }
  for new_feature in remaining_features
    current_feature_set := current_best_feature_set ∪ {new_feature}
    error(new_feature):= classific_test(current_feature_set)
  end
  best_feature := arg min error
  current_best_feature_set := current_best_feature_set ∪ {best_feature}
  remaining_features := remaining_features \ {best_feature}
end
feature_ranking : = current_best_feature_set
```

Listing 1: Pseudo-code for the Sequential Forward Feature Selection method.

Figure 3 shows the classification error as a function of the feature set size. For this procedure the Support Vector Regression estimation method was applied, which is introduced in Section 4.1. With an increasing number of features the classification error shrinks rapidly. However, for large feature sets, the error increases again, though only marginally. This effect might be caused by the fact that some features bear contradictory information concerning the emotion that is being conveyed. Thus, we found that, for each of the primitives and each of the classifiers, using 20 features was sufficient. Adding more features did not improve the results. The variance of the error was almost constant for feature sets of 10 or more features with a value of $\sigma^2 \approx 0.015$.

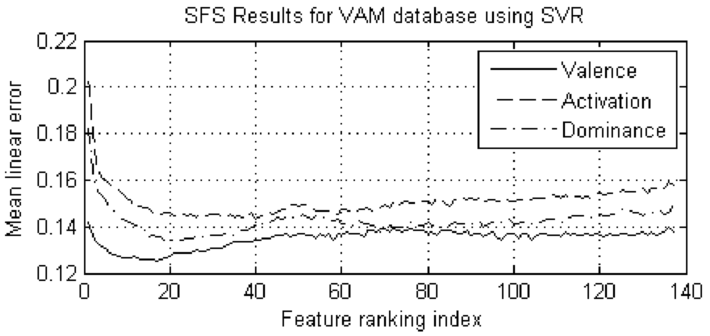

Figure 3: Classification error as a function of the feature set.

The SFS method was compared to Principal Component Analysis (PCA). While SFS includes the classifier in the selection routine, the PCA does not use the classification result as a feedback. It is based only on the N = 893 observations of the M = 137 features,

$$\mathbf{v}_n = (v_{n,1}, \dots, v_{n,M})^{\mathrm{T}}, \qquad n = 1, \dots, N, \tag{1}$$

which are combined in the observation matrix

$$\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_N). \tag{2}$$

Note that the features have to be zero-mean, which can be achieved by subtracting the estimated mean value $\bar{v}_m = \frac{1}{N}\sum_{n=1}^{N} v_{n,m}$ for each feature $m = 1,\dots,M$. The $M \times M$ covariance matrix of the features can thus be stated as $\mathbf{C}_{\mathbf{VV}} = \mathbf{VV}^{\mathrm{T}}$. According to (Kroschel, 2004), the features can be decorrelated by transforming them to a orthonormal basis $\mathbf{\Phi} = (\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_M)$, where the basis vectors $\boldsymbol{\varphi}_m$, $m = 1,\dots,M$, are determined by solving the deterministic equation

$$\mathbf{C}_{\mathbf{VV}} \cdot \boldsymbol{\varphi}_m = \lambda_m \cdot \boldsymbol{\varphi}_m. \tag{3}$$

Thus, the basis vectors are the eigenvectors of the covariance matrix $\mathbf{C}_{\mathbf{VV}}$. The transformed, uncorrelated features $\mathbf{u}_n$ can be calculated by

$$\mathbf{u}_n = \mathbf{\Phi}^* \mathbf{v}_n, \tag{4}$$

where $\mathbf{\Phi}^*$denotes the complex conjugate transpose of $\mathbf{\Phi}$. In the PCA, now the eigenvectors are sorted in decreasing order of the eigenvalues, and only those eigenvectors are kept as new basis vectors whose eigenvalue exceeds a threshold of, in our case, 1% of the maximum eigenvalue:

$$\tilde{\mathbf{u}}_n = \tilde{\mathbf{\Phi}}^* \mathbf{v}_n, \tag{5}$$

with

$$\tilde{\mathbf{\Phi}} = (\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_{\tilde{M}}), \quad \tilde{M} < M. \tag{6}$$

Figure 4 shows the eigenvalues of the covariance matrix based on our observations of acoustic features derived from emotional speech. It can be seen that the eigenvalues decrease quickly with increasing index. Thus, the first ten components carry already a large amount of the information. However, to include 90% of the total eigenvalue sum, at least $\tilde{M} = 61$ components must be included in the uncorrelated feature set; to include 99%, $\tilde{M} = 96$ components are required.

Since such a large number of features is not desirable for the classifier, we preferred the SFS method to the PCA method in order to reduce the size of the feature set. Note that SFS also reduces the computational demand since only a smaller number of features have to be calculated in contrast to PCA where still all features would be necessary to apply Equation (5).

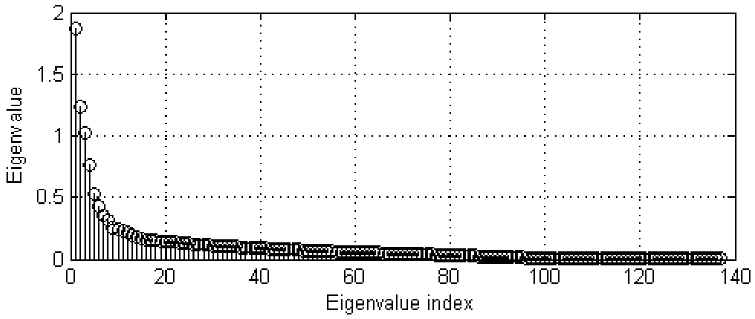

Figure 4: Eigenvalues of the covariance matrix derived from 893 observations of 137 acoustic features extracted from emotional speech.

4. Emotion primitives estimation

The task of the emotion estimator is to map the acoustic features to the real-valued emotion primitives. We analyzed several classifiers: Support Vector Regression, Fuzzy *k*-Nearest Neighbor classifiers, and a rule-based Fuzzy Logic inference method. In the following subsections we briefly describe the individual estimators. The desired output is not a classification into one of a finite set of categories but an estimation of continuous-valued emotion parameters, the primitives $x_n^{(i)} \in [-1,+1] \subset \mathbb{R}$, $i \in \{valence,\ activation,\ dominance\}$. The results of these three estimators are discussed in Section 5.

4.1. Support Vector Regression

Support Vector Regression (SVR) is a regression method based on Support Vector Machines (Vapnik, 1995; Campbell, 2001; Schölkopf & Smola, 2001). Support Vector Machines are applied to a wide range of classification tasks (Abe, 2005). Based on a solid theoretical framework, they were shown to not only minimize the empirical training error but, more general, the structural risk. The decision function in SVMs is found by the hyperplane which maximizes the margin between two classes. Non-linear classification can be applied very efficiently by using the so-called *kernel trick*, i.e., by replacing the inner products that appear in the calculation of the decision function by (optionally) non-linear kernel functions. This kernel trick replaces a transformation of the features into a higher-dimensional space, linear classification in this higher-dimensional space, and re-transformation into the original space.

In Support Vector Regression the role of the separation margin is inverted, i.e., the aim is to find the optimal regression hyperplane so that most training samples lie *within* an ε-margin around this hyperplane. Figure 5 shows a schematic sketch for SVR in which the non-linear regression curve was found using the kernel trick. The technical terms included in the figure will be described below. First results of using SVR for emotion primitives estimation were reported in (Grimm et al., 2007b).

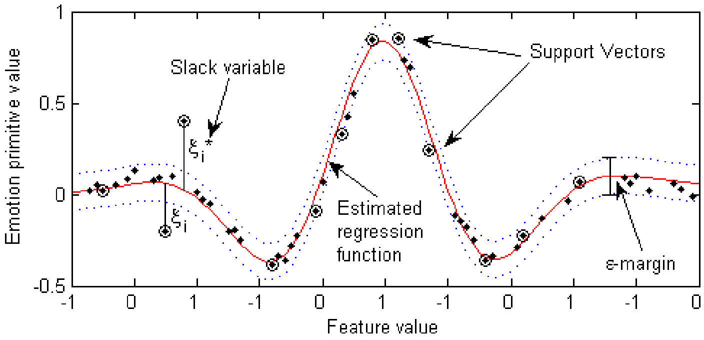

Figure 5: An example for Support Vector Regression using the kernel trick.

The mathematical formulation of the problem how to find an optimal regression hyperplane using a finite set of training samples can be found in (Vapnik, 1995; Schölkopf & Smola, 2001). Here we give only a short summary.

The goal is to find a function $f^{(i)}$ which maps the m acoustic features $\mathbf{v} = (v_1, \dots, v_m)^{\mathrm{T}} \in \mathbb{R}^m$ to the emotion primitive value $x^{(i)} \in \mathbb{R}$,

$$\hat{x}^{(i)} = f^{(i)}(\mathbf{v}). \tag{7}$$

We need three SVR functions to estimate the three emotion primitives separately: $i \in \{valence, activation, dominance\}$. For improved readability we omit the notion of the index (i) in the following.

The easiest solution is a linear function, i.e., a hyperplane in $\mathbb{R}^m$,

$$f(\mathbf{v}) = \langle \mathbf{w}, \mathbf{v} \rangle + b, \tag{8}$$

where $\mathbf{w} \in \mathbb{R}^m$ and $b \in \mathbb{R}$ are the parameters of the hyperplane and $\langle \cdot \rangle$ denotes the inner product. To determine the parameters we need a set of N learning samples $(\mathbf{v}_n;\ x_n), n = 1, \dots, N,$ and a loss function $l(\eta)$ to penalize the distance between the function's output $\hat{x}_n$ and the (for the training samples well-known) true x_n,

$$l(\eta) = l(\hat{x}_n - x_n). \tag{9}$$

We chose the ε-insensitive loss function

$$l_\varepsilon(\eta) = \begin{cases} -\eta - \varepsilon & \text{for} \quad \eta < -\varepsilon \\ 0 & \text{for} \quad -\varepsilon \le \eta \le \varepsilon, \\ \eta - \varepsilon & \text{for} \quad \eta > \varepsilon \end{cases} \tag{10}$$

which assigns zero loss within a margin of width ε around the true value and linear loss for larger deviations. This loss function allows for the neglection of a large amount of training samples in the calculation of the hyperplane (Smola, 1996). The remaining samples are called *Support Vectors* (see Figure 5).

Since the structural risk is minimized for the function f with the least complexity, the problem can be formulated as

$$\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to } \begin{cases} x_n - (\langle \mathbf{w}, \mathbf{v}_n \rangle + b) \le \varepsilon \\ (\langle \mathbf{w}, \mathbf{v}_n \rangle + b) - x_n \le \varepsilon \end{cases} \quad \text{for } n = 1,\ldots,N. \tag{11}$$

It is common and in our applications absolutely necessary to allow some outliers. This can be achieved by introducing *slack variables* ξ_n, ξ_n^* (see Figure 5) and a *soft margin parameter* C, which yields the following problem:

$$\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}(\xi_n + \xi_n^*)$$

$$\text{subject to } \begin{cases} x_n - (\langle \mathbf{w}, \mathbf{v}_n \rangle + b) \le \varepsilon + \xi_n^* \\ (\langle \mathbf{w}, \mathbf{v}_n \rangle + b) - x_n \le \varepsilon + \xi_n \\ \xi_n, \xi_n^* \ge 0 \end{cases} \quad \text{for } n = 1,\ldots,N. \tag{12}$$

This problem can be reformulated using the dual Lagrange function involving the Lagrange multipliers α_n, α_n^* as functions of the slack variables,

$$\text{maximize } -\frac{1}{2}\sum_{n,l=1}^{N}(\alpha_n^* - \alpha_n)(\alpha_l^* - \alpha_l)\langle \mathbf{v}_n, \mathbf{v}_l \rangle - \varepsilon\sum_{n=1}^{N}(\alpha_n^* + \alpha_n) + \sum_{n=1}^{N}x_n(\alpha_n^* - \alpha_n)$$

$$\text{subject to } \begin{cases} \sum_{n=1}^{N}(\alpha_n^* - \alpha_n) = 0 \\ \alpha_n, \alpha_n^* \in \left[0, \frac{C}{N}\right]. \end{cases} \tag{13}$$

In this maximization task the training samples $\mathbf{v}_n$ only occur as inner products. Since one of the Lagrange conditions requires

$$\mathbf{w} = \sum_{n=1}^{N}(\alpha_n^* - \alpha_n)\mathbf{v}_n, \tag{14}$$

which means that $\mathbf{w}$ can completely be expressed as a linear combination of the feature vectors (*Support Vector Expansion*), the target function f can be stated as

$$f(\mathbf{v}) = \sum_{n=1}^{N} (\alpha_n^* - \alpha_n) \langle \mathbf{v}_n, \mathbf{v} \rangle + b. \tag{15}$$

While (15) is a functional formulation, it is easily applied to any query feature vector **v**0 with unknown emotion primitive values $x^{(i)}$ by evaluating $f(\mathbf{v} = \mathbf{v}_0)$.

It is obvious that both in (13) and (15) the feature vectors only occur as inner products. These inner products can be replaced by a (non-linear) kernel function:

$$\langle \mathbf{v}_n, \mathbf{v}_l \rangle \to K(\mathbf{v}_n, \mathbf{v}_l). \tag{16}$$

This replacement expresses the kernel trick mathematically. It allows non-linear regression in an efficient way.
We used the following kernel functions:

- Radial basis function (SVR-RBF):

$$K(\mathbf{v}_n, \mathbf{v}_l) = \exp\{-\|\mathbf{v}_n - \mathbf{v}_l\|^2/(2\sigma^2)\} \tag{17}$$

- Polynomial kernel (SVR-Poly):

$$K(\mathbf{v}_n, \mathbf{v}_l) = (\langle \mathbf{v}_n, \mathbf{v}_l \rangle + 1)^d \tag{18}$$

Both, the choice of the kernel function and the kernel parameter, is important. If σ in SVR-RBF is small, the regression function will follow closely the training samples. For σ = 0.01, e.g., we can observe clear overfitting. The larger σ, the flatter gets the target curve; therefore it cannot be chosen too big. We tested for the range of $\sigma \in [10^{-4}, 10^{4}]$ and found that $\sigma \in [2.5, 10]$ gives good results. We finally chose $\sigma = 3.5$.
In the polynomial kernel the parameter d determines the order of the polynomial. The higher d, the more complex gets the regression curve. However, we found that, on a search interval of $d \in [1, 10]$, the best results can be achieved with $d = 1$. This choice results in a polynomial of order 1, i.e., a linear function.

In addition to the choice of the kernel function, the parameters of the SVR, C and ε, have to be set. We used a first, grid-based search on a logarithmic scale and a second, fine-grained search in the best region to find those parameters that give the lowest error (Chudoba, 2006). The soft margin parameter C has an influence on how many sample vectors are finally used for the calculation of the regression curve. The higher C, the more outliers are allowed, and the more sample vectors are included as support vectors. Our search region was $C \in [10^{-4}, 10^{4}]$, and the best results were achieved for $C \in [0.1, 50]$. Since this parameter has to be regarded jointly with the kernel function parameter, we chose $C = 10$ for SVR-RBF and $C = 0.1$ for SVR-Poly, respectively.

The parameter ε defines the width of the margin around the regression curve. This parameter essentially influences the number of support vectors used. Since, in general, those feature vectors are used as support vectors which are lying outside or at least close to the margin border, a larger margin yields less support vectors. The number of support vectors is computationally relevant because the summation in (15) in practice reduces to the summation of the support vectors only. Testing on the range of $\varepsilon \in [0,1]$, we decided to choose $\varepsilon = 0.2$ for all kernels.

For all experiments, the *libsvm* implementation was used (Chang & Lin, 2001).

4.2. Fuzzy k-Nearest Neighbor estimator

As an alternative to SVM, the *k*-Nearest Neighbor method was studied. This method was shown to give comparable results on related problems, namely the discrete categorization of emotion stereotypes (Yacoub et al., 2003; Dellaert et al., 1996). The *k*-Nearest Neighbor (KNN) method is a distance-based approach. It determines the *k* closest neighbors of a query $\mathbf{v} = \mathbf{v}_0$ in the feature space and assigns the properties of these neighbors to the query (Kroschel, 2004). Such method can be regarded as spanning a hypersphere *S* around the feature vector $\mathbf{v}_{0,}$

$$S = \{\mathbf{v} \in V \,|\, \|\mathbf{v} - \mathbf{v}_0\|_p \leq r\}, \tag{19}$$

where

$$\mathrm{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_N\} \subset \mathbb{R}^M \tag{20}$$

is the set of all sample vecors $\mathbf{v}_n$, $n = 1,\dots,N$, and the radius *r* is chosen to have exactly *k* sample feature vectors lie within this hypersphere,

$$|S| = k. \tag{21}$$

Thus, without loss of generality, *S* can be stated as

$$S = \{\mathbf{v}_{n_1}, \dots, \mathbf{v}_{n_k}\}, \tag{22}$$

with the indices $n_1, \dots, n_k \in [1, \dots, N]$.

In our case, the properties of the neighbors are the emotion primitive values $x_{n_1}^{(i)}, \dots, x_{n_k}^{(i)}$. These values are averaged to get the final emotion estimate $\hat{x}^{(i)}$ for a query feature vector $\mathbf{v}_0$,

$$\hat{x}^{(i)} = \frac{1}{k} \sum_{\kappa=1}^{k} x_{n_\kappa}^{(i)}. \tag{23}$$

Due to this average, all *k* neighbors have an influence on the estimate. This led us to calling the method *Fuzzy KNN.*

The parameters to choose are p of the L_p distance in (19) and k, the number of neighbors considered. According to (Kroschel, 2004), the vector distance $\|\cdot\|_p$ in (19) is defined as

$$\|\mathbf{v}-\mathbf{v}_0\|_p = \left(\sum_{m=1}^{M} |v_m - v_{0,m}|^p\right)^{\frac{1}{p}}. \tag{24}$$

We tested $p \in \{1, 1.5, \ldots, 4\}$ and found that the results were almost independent of the distance norm. Thus, we finally chose $p = 2$ (Euclidean distance).

The second design parameter, k, had a greater impact on the results. We tested $k \in \{1, 3, \ldots, 15\}$ and found that the error decreased rapidly with increasing k. The error remained at a relatively constant level for values of $k \geq 9$. Thus, we decided for $k = 11$ for our experiments.

4.3. Rule-based Fuzzy Logic estimator

A rule-based Fuzzy Logic (FL) estimator has previously been used for automatic emotion primitive estimation (Grimm & Kroschel, 2005b; Grimm et al., 2007a). Therefore, it will be described very briefly here. This method can be regarded as the state-of-the art in emotion primitive estimation.

The fuzzy logic captures well the nature of emotions, which in general is fuzzy in description and notation. This fuzzy description is reflected in the number of linguistic terms which are used to describe feelings, moods, and affective attitudes.

A fuzzy logic estimator consists of three major elements (Kroschel, 2004):

- Fuzzification
- Inference
- Defuzzification

The fuzzification step transforms the crisp variables, which are the acoustic features in our case, into fuzzy, linguistic variables. We transformed each feature into three linguistic variables, *low, medium,* and *high,* respectively. We assigned membership grades to each of these fuzzy variables according to the relative position of the crisp feature value within the range between the 10% and the 90% quantiles of the overall feature value distribution observed in the training samples.

The rules in the inference system can be derived from expert knowledge. However, we decided to derive them automatically by analysis of the relation between the acoustic features and the desired emotion primitives. We used a set of three linguistic, fuzzy variables for each primitive: *negative, neutral,* and *positive* for *valence*; *calm, neutral,* and *excited* for *activation*; and *weak, neutral,* and *strong* for *dominance*. Thus, each fuzzy variable of the acoustic features was related to each fuzzy variable of the emotion primitives.

The individual steps of the inference part are the following (Kroschel, 2004; Grimm et al., 2007a): First, all features were aggregated using maximum aggregation method. Then, inference was performed using the product method; this process determines the conclusion drawn from the fuzzy acoustic features onto the fuzzy emotion values. Finally, accumulation of the three fuzzy membership contours for each primitive was achieved using the maximum method. The result was the fusion of *negative, neutral,* and *positive* contours into one membership function for *valence,* etc.

In the last step, the fuzzy emotion values were defuzzified to yield crisp emotion primitive values. We used the centroid method for this task.

5. Results

This section reports the results of our experiments on estimating the emotion primitives conveyed in spontaneous speech. First, the results are compared with respect to the different estimators. While these results were already presented very briefly in (Grimm et al., 2007b), we provide a more elaborate discussion of these results here. Second, we present the learning curves of the different estimators to show the dependence on the data size that is necessary for training.

5.1 Comparison of the emotion estimation results

All experiments were performed using a 10-fold cross-validation. To assess the estimation results we used two different measures:

- The mean linear error between the emotion estimates $\hat{x}_n^{(i)}$ and the reference annotated manually by the human evaluators, $x_n^{(i)}$,

$$e^{(i)} = \sum_{n=1}^{N} \left| \hat{x}_n^{(i)} - x_n^{(i)} \right| \qquad (25)$$

- The empirical correlation coefficient between the emotion estimates and the reference,

$$r^{(i)} = \frac{\sum_{n=1}^{N} \left(\hat{x}_n^{(i)} - \bar{\hat{x}}^{(i)} \right) \left(x_n^{(i)} - \bar{x}^{(i)} \right)}{\sqrt{\sum_{n=1}^{N} \left(\hat{x}_n^{(i)} - \bar{\hat{x}}^{(i)} \right)^2 \sum_{n=1}^{N} \left(x_n^{(i)} - \bar{x}^{(i)} \right)^2}} \qquad (26)$$

Table 1 summarizes (a) the mean linear error for each estimator, for each emotion primitive separately, and (b) the correlation coefficient to measure the tendency in the emotion estimates.

The results in Table 1(a) indicate that all primitives can be estimated with a small error in the range of 0.13 to 0.18. There was only one exception when *valence* was estimated using the FL estimator (0.27). Considering the range of values, which was [-1,+1], it can be stated that the emotion primitives are mostly estimated in the correct region, and, e.g., a very excited utterance might get estimated as moderately excited to very excited, but not as very calm.

	(a) Mean Error			(b) Correlation Coefficient		
	Valence	*Activation*	*Dominance*	*Valence*	*Activation*	*Dominance*
SVR-RBF	0.13	0.15	0.14	0.46	0.82	0.79
SVR-Pol	0.14	0.16	0.15	0.39	0.80	0.77
FL	0.27	0.17	0.18	0.28	0.75	0.72
KNN	0.13	0.16	0.14	0.46	0.80	0.78

Table 1: Mean error and correlation with reference of the emotion primitive estimation.

The error is thus even better than the human evaluation, which showed on average a standard deviation of 0.31 (c.f. Section 2.2). However, it has to be noted that this standard deviation also includes the quantization error of the emotion axes due to the finite set of values offered to the evaluators. Also the automatic estimation might benefit from the relatively large amount of neutral and moderate emotions that yielded better individual results than the extreme emotions.

The correlation between the estimates and the reference was significantly different for the individual emotion primitives, as shown in Table 1(b). The correlation for *valence* was between 0.28 and 0.46 for the different estimators, and it was between 0.72 and 0.82 for *activation* and *dominance*. It has to be noted that the correlation coefficients for *valence* are only moderately significant at $p > 10^{-3}$, while all other correlation coefficients are statistically significant at $p < 10^{-5}$. These different significance levels reflect the fact that the distribution in the values for *valence* was much narrower than for *activation* or *dominance*. Thus the results imply very good recognition results for *activation* and *dominance*, and moderate recognition results for *valence*.

Comparing the individual estimation methods, we can say that the best results were achieved using the SVR-RBF estimator with errors of 0.13, 0.15, and 0.14, and correlation coefficients of 0.46, 0.82, and 0.79 for *valence*, *activation*, and *dominance*, respectively. The Fuzzy KNN estimator performed almost as well as the SVR-RBF. The SVR-Poly estimator gave worse results for *valence* in comparison to almost as good results for *activation* and *dominance*. Similarly, the FL estimator gave even worse results for *valence* but still very good ones for *activation* and *dominance*.

Thus, the kernel-based method outperformed all other methods only marginally, apart from the FL method which seems to be clearly the last choice. These differences are probably caused by the complexity in the representation of the relation between the acoustic features and the emotion primitives. While Fuzzy Logic only uses the rules derived from the correlation coefficients, which is a rather coarse generalization, the SVR method provides a more complex, nonlinear relation in form of the regression curve in a high-dimensional feature space. KNN does not provide such sophisticated abstraction, but uses a huge amount of comparison options to choose the neighbors from.

Thus, on deciding between SVR and KNN, the focus must be set on the computational demand: If we have enough computational power to calculate all the distances necessary in the KNN method, this might be the choice due to the simplicity in the setup of the classifier. Alternatively, if we are forced to have low computational demands at runtime, we might decide for SVR-RBF and provide the computational power only once, which is at the instant of generating the regression hyperplane.

As a summary of the results, Figure 6 shows the estimation results using the SVR-RBF estimator. In this graph, the emotion primitive values are arranged in ascending order. This order is in contrast to the experiment, where we provided random order, but helps in grasping the benefits and the limits of the presented method, in particular with respect to the mentioned difficulty of having relatively few positive emotions but a wide range of *activation* and *dominance* values.

Figure 6 reveals some information about the nature of the errors. Most of the estimates are located within a small margin around the references. However, a small number of very high or very low primitive values was occasionally underestimated.

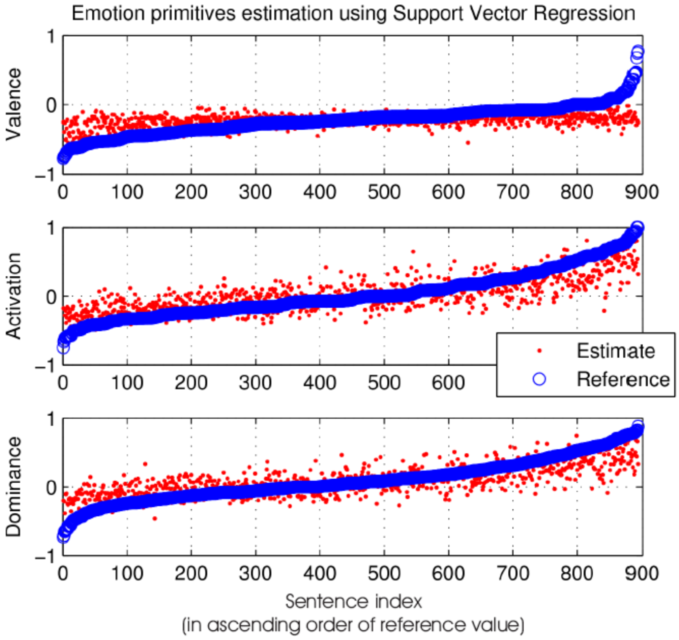

Figure 6: Emotion primitives estimation: results in comparison with manually labeled human reference.

5.2 Learning curves

In addition to the direct comparison of the estimator results, it is interesting to see the amount of training data that is necessary to achieve such results. We analyzed a wide range of 50% to 98% of the data to be available for training and compared the estimation results for these individual conditions. Figure 7 shows these learning curves for (a) SVR-RBF, (b) FL, and (c) KNN.

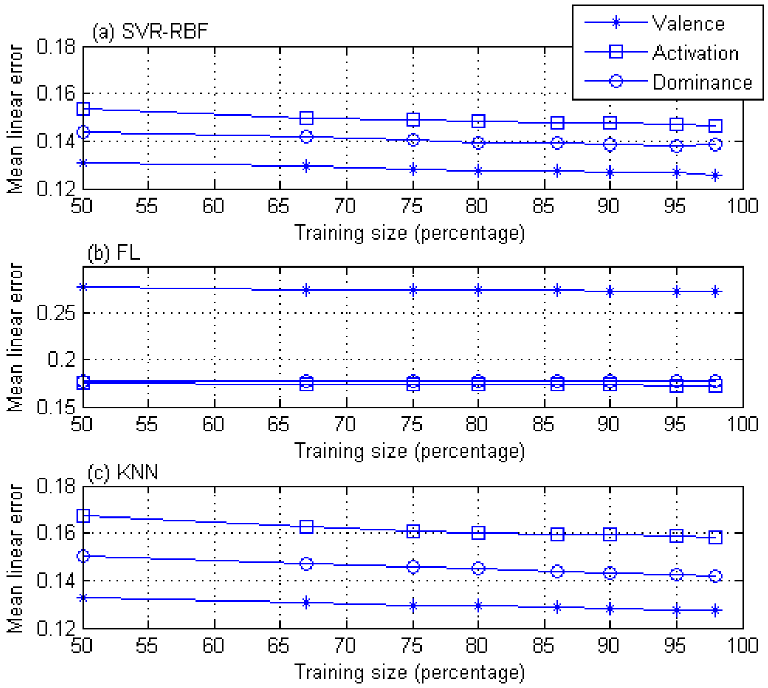

Figure 7: Learning curves of different estimators for emotion primitive estimation, based on the VAM corpus.

It can be seen that the error curves are very different. For the FL estimator, the error curves are almost constant over the total range. This result is very interesting since it shows that although the FL absolute error was higher than the one using alternative estimators, it is more robust considering the necessary amount of training data. Therefore this method might be worth using for applications in which few training data is available.

If we compare the SVR-RBF and the KNN learning curves, we can observe that the error for KNN depends more on the training data size than the respective error using SVR. This difference can be explained by the nature of learning in these two methods. SVR uses a generalizing method that yields a more abstract representation of the training data in form of the regression curve parameters. In contrast, KNN uses only an explicit representation of a set of training features. Thus, providing a smaller number of training samples directly reduces the options for comparison when trying to find the most appropriate neighbors.

6. Conclusion

In this chapter we discussed the recognition of emotions in spontaneous speech. We used a general framework motivated by emotion psychology to describe emotions by means of three emotion "primitives" (attributes), namely *valence*, *activation*, and *dominance*. With these emotion primitives, we proposed a real-valued three-dimensional emotion space concept to overcome the limitations in the state-of-the-art emotion categorization. We tested the method on the basis of 893 spontaneous emotional utterances recorded on a German TV talk-show.

For the acoustic representation of the emotion conveyed in the speech signal, we extracted 137 features. These reflected the prosody and the spectral characteristics of the speech. We tested two methods to reduce the problem of large feature sets, Principal Component Analysis and Sequential Feature Selection. Thus, we selected the 20 most relevant acoustic features that yielded the best recognition results.

For the estimation of the emotion primitives, Support Vector Regression, Fuzzy Logic, and Fuzzy k-Nearest Neighbor methods were used. We found that the emotion primitives could be estimated with a small error of 0.13 to 0.15, where the range of values was [-1,+1]. The correlation between the reference annotated manually by the evaluators and the automatically calculated estimates was moderate (0.46, *valence*) to high (0.82/0.79, *activation*/*dominance*). In comparison to the Fuzzy Logic estimator, which was the baseline, the error for *valence, activation* and *dominance* estimation could be reduced by 52%, 12% and 22%, respectively.

Thus, Support Vector Regression gave the best estimation results, however, closely followed by KNN. Note that while SVR is computationally much more demanding for initialization (finding the regression hyperplane), the KNN method requires more computational power at the actual estimation step due to the distance matrix that has to be calculated. The rule-based FL algorithm is computationally less demanding but gives clearly inferior results, at least for *valence*. However, when regarding the learning curves of the three estimators, i.e., assessing the estimation error as a function of the training data size, it was shown that the Fuzzy Logic method gave the most robust results. Thus, in the case of very few training data available, the FL method might be an appropriate choice again.

In our future work we will study the fusion of the emotion primitive estimation with the automated speech recognition (ASR). While the emotion recognition might be used to parameterize or personalize the ASR unit, the phoneme estimates of the ASR, on return, might be used to improve the emotion recognition. Future work will also investigate the design of a real-time system using the algorithms that were reported here. The advantage of continuous-valued estimates of the emotional state of a person could be used to build an adaptive emotion tracking system. Such a system might be capable to adapt to individual personalities and long-term moods, and thus finally provide indeed humanoid man-machine interfaces.

7. Acknowledgment

This work was supported by grants of the Collaborative Research Center (SFB) 588 "Humanoid Robots – Learning and Cooperating Multimodal Robots" of the Deutsche Forschungsgemeinschaft (DFG). We thank Robert Chudoba who contributed a relevant part to this research in his Master Thesis.

8. References

Abe, S. (2005). *Support Vector Machines for Pattern Recognition.* Berlin, Germany: Springer.

Campbell, C. (2001). An Introduction to Kernel Methods. In R. Howlett, & L. Jain (Eds.), *Radial Basis Function Networks 1* (pp. 155-192). Heidelberg, Germany: Physica-Verlag.

Chang, C.-C., & Lin, C.-J. (2001). *LIBSVM: A Library for Support Vector Machines.* Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chudoba, R. (2006). *Klassifikation von Emotionen mit Support Vector Machines.* Karlsruhe, Germany: Universität Karlsruhe (TH): Diploma Thesis.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine , 18* (1), pp. 32-80.

Cowie, R., & Cornelius, R. (2003). Describing the Emotional States That Are Expressed in Speech. *Speech Communication , 40*, pp. 5-32.

Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing Emotion in Speech. *Proc. International Conference on Spoken Language Processing, 3*, pp. 1970-1973. Philadelphia, PA, USA.

Fischer, L., Brauns, D., & Belschak, F. (2002). *Zur Messung von Emotionen in der angewandten Forschung.* Lengerich, Germany: Pabst Science Publishers.

Fragopanagos, N., & Taylor, J. (2005). Emotion Recognition in Human-Computer Interaction. *Neural Networks , 18* (4), pp. 389-405.

Grimm, M., & Kroschel, K. (2005a). Evaluation of Natural Emotions Using Self Assessment Manikins. *Proc. ASRU*, (pp. 381-385).

Grimm, M., & Kroschel, K. (2005b). Rule-Based Emotion Classification Using Acoustic Features. *Proc. Int. Conf. on Telemedicine and Multimedia Communication.*

Grimm, M., Mower, E., Kroschel, K., & Narayanan, S. (2006). Combining Categorical and Primitives-Based Emotion Recognition. *Proceedings European Signal Processing Conference (Eusipco).* Florence, Italy.

Grimm, M., Mower, E., Kroschel, K., & Narayanan, S. (2007a). Primitives-Based Evaluation and Estimation of Emotions in Speech. *Speech Communication* .

Grimm, M., Kroschel, K., & Narayanan, S. (2007b). Support Vector Regression for Automatic Recognition of Spontaneous Emotions in Speech. *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Honolulu, HI, USA: Accepted for Publication.

Kehrein, R. (2002). The prosody of authentic emotions. *Proc. Speech Prosody Conf.*, (pp. 423-426).

Kittler, J. (1978). Feature Set Search Algorithms. *Pattern Recognition and Signal Processing* , pp. 41-60.

Kroschel, K. (2004). *Statistische Informationstheorie* (4. ed.). Berlin, Germany: Springer.

Lee, C. M., Narayanan, S., & Pieraccini, R. (2001). Recognition of Negative Emotions from the Speech Signal. *Proc. IEEE Automatic Speech Recognition and Understanding Wsh. (ASRU).*

Lee, C. M., & Narayanan, S. (2005). Toward Detecting Emotions in Spoken Dialogs. *IEEE Transactions on Speech and Audio Processing , 13* (2), pp. 293-303.

Murray, I., & Arnott, J. (1993). Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal of the Acoustic Society of America , 93* (2), pp. 1097-1108.

O'Shaughnessy, D. (1999). *Speech Communications: Human and Machine* (2. ed.). John Wiley & Sons Inc.

Picard, R. (1997). *Affective Computing.* Cambridge, MA, USA: MIT Press.

Scherer, K. (1990). *Psychologie der Emotion.* Göttingen, Germany: Hogrefe.

Schölkopf, B., & Smola, A. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA, USA: The MIT Press.

Schröder, M. (2003). *Speech and Emotion Research: An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis.* Universität des Saarlandes, Germany: Ph.D. Thesis.

Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., & Rigoll, G. (2005). Speaker Independent Speech Emotion Recognition by Ensemble Classification. *Proc. Int. Conf. on Multimedia and Expo*, (pp. 864-867).

Smola, A. (1996). *Regression Estimation with Support Vector Learning Machines.* Master Thesis: Technische Universität München.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer.

Vidrascu, L., & Devillers, L. (2005). Real-Life Emotion Representation and Detection in Call Centers Data. *Proc. Int. Conf. on Affective Computing and Intelligent Interaction*, (pp. 739-746).

Yacoub, S., Simske, S., Lin, X., & Burns, J. (2003). *Recognition of Emotions in Interactive Voice Response Systems.* Palo Alto, USA: HP Laboratories .

Yu, C., Aoki, P., & Woodruff, A. (2004). Detecting User Engagement in Everyday Conversations. *Proc. Int. Conf. Spoken Lang. Processing*, (pp. 1329-1332).