FUSING ACOUSTIC AND LINGUISTIC PARAMETERS

FOR

MULTILINGUAL EMOTION RECOGNITION

by

Mustafa Erden

B.S., Electrical And Electronics Engineering, Boğaziçi University, 2008

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Department of Electrical and Electronics Engineering
Boğaziçi University

2011

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge my thesis supervisor Prof. Levent Arslan for his kindness and understanding. His guidance was invaluable to me at all stages of preparation of this thesis.

I am grateful to all my friends from Sestek for their friendship and their help with data labeling. I would like particularly to acknowledge Ali Haznedaroğlu and Osman Büyük for technical support and constructive comments.

I would like to thank to University of Erlangen and Sestek A.Ş. for kindly providing the databases that are investigated.

Finally I would like to thank my family for their love, support and encouragement.

ABSTRACT

FUSING ACOUSTIC AND LINGUISTIC PARAMETERS FOR MULTILINGUAL EMOTION RECOGNITION

Emotion recognition from speech can be used for detection of customer problems in call centers, agent performance monitoring, improving automatic speech recognition accuracies, enhancing human robot as well as human machine interaction. In this thesis two different spontaneous databases are investigated in terms of binary emotion classification. On Turkish call center dataset (CCD) which consists of human-human dialogs, emotion recognition problem is defined on angry and non-angry classes. On Fau Aibo dataset (FAD) which is composed of recordings of children playing with a pet robot, the negative and idle classes are considered.

For extracting acoustic information we have implemented Support Vector Machines with utterance level features and Gaussian Mixture Models with frame level features. In terms of language modeling we compared word based, stem-only and stem+ending structures using manual transcriptions. Stem+ending based system resulted in the highest accuracies on CCD whereas the word based LM performed the best on FAD. This can be mainly attributed to the agglutinative nature of Turkish language. When we fused the acoustic and LM classifiers using a Multi Layer Perceptron (MLP) we could achieve 89% and 69% correct detection of both classes for CCD and FAD respectively.

ÖZET

AKUSTİK VE DİLBİLİMSEL PARAMETRELERLE ÇOK DİLLİ DUYGU TANIMA

Sesten duygu tanıma çağrı merkezlerinde problemlerin tespitinde, operatör performansının gözlemlenmesinde, otomatik konuşma tanıma sistemlerinin iyileştirilmesinde, insan-robot ve insan-makine etkileşiminin artırılmasında kullanılabilir. Bu tezde iki farklı doğal veritabanı ikili duygu tanıma açısından incelenmiştir. Türkçe çağrı merkezi veritabanı üzerinde duygu tanıma problemi sinirli ve sinirli-olmayan sınıflar üzerinde tanımlanmıştır. Robot bir evcil hayvanla oynayan çocukların kayıtlarından oluşan Almanca Fau Aibo veritabanı üzerinde de negatif ve negatif-olmayan sınıflar göz önünde bulundurulmuştur.

Akustik bilgiyi çıkarmak için sözcük grubu bazında parametreler kullanılarak Destekçi Vektör Makineleri ve çerçeve bazında parametreler kullanılarak Gauss Karışım Modelleri uygulanmıştır. Dil modeli için de kayıtların manuel transkripsiyonları kullanılarak oluşturulan kelime, sadece-kök ve kök+ek tabanlı modeller karşılaştırılmışlardır. Çağrı merkezi veritabanında kök+ek bazlı dil modeli en yüksek sonuçlar sağlarken, Fau Aibo veritabanında kelime tabanlı dil modeli en iyi sonuçları vermiştir. Bu durum Türkçenin bitişken yapısına bağlanabilir. Akustik ve dil modeli sınıflandırıcılarının skorları çok katmanlı algılayıcı kullanılarak birleştirildiğinde, çağrı merkezi ve Fau Aibo veritabanları için sırasıyla 89% ve 69% doğru tanıma elde edilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

b	Hyperplane shift
c_i	i'th coefficient
d	Vector dimension
f	Frequency
$K(.,.)$	Kernel function
m_j	Log filter bank amplitudes
N	Number of frequency bands
N	Number of Gaussians
$\vec{x_i}$	Data vector
y_i	Class of i'th sample
V	Total number of observed words
$\vec{w}$	Normal vector to separating hyperplane
w_i	i'th word
$\vec{\mu}_i$	Mean
$\mathbf{\Sigma}_i$	Covariance matrix
ξ_i	Slack variable

LIST OF ACRONYMS/ABBREVIATIONS

ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
CCD	Call Center Dataset
DTMF	Dual-Tone Multi-Frequency
FAD	Fau Aibo Dataset
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IVR	Integrated Voice Response
LLD	Low Level Descriptors
LM	Language Model
MFCC	Mel Frequency Cepstral Coefficients
MPL	Multi Layer Perceptron
NN	Neural Network
OOV	Out of Vocabulary
RAPT	Robust Algorithm for Pitch Tracking
SVM	Support Vector Machines
UA	Unweighted Average
WOZ	Wizard of Oz

1. INTRODUCTION

Speaking is the natural way of human communication. Apart from semantic content, speech contains multiple clues about the speaker. While the unique speaker characteristics can be utilized to verify the identity of the person, the accent reflects the region where he or she is from.

Emotional state of the speaker is another information that can be extracted from speech. This can be used in various applications such as prioritizing voice mails, enriching multimedia content, enhancing human robot interaction. Also in integrated voice response systems (IVR) by monitoring the emotion of the customer, adaptive scenarios can be used or the call can be transferred to a human operator when a disturbance is sensed.

In automatic speech recognition systems performance degrades with the presence of emotions. Using emotion detection and decoding with specific acoustic models adapted to different emotions can enhance the ASR performance significantly [1].

In call centers all dialogues between agents and clients are recorded for quality measurement and agent performance monitoring. However, because of additional costs and large amounts of accumulated data, only a small fraction of those conversations are reviewed. If automatic emotion recognition is applied to this data problematic calls can be identified to some extent and retention operations can be performed for disgruntled customers.

Emotion recognition on real life call center data is a challenging task. First of all, the environmental conditions and microphone qualities are from a broad range. Also it is impossible to know the exact emotion of the speaker. Therefore subjective tests are necessary for labeling the data. During these tests not always the subjects agree on a common label. In [2], an entropy based method is used to evaluate the classification results. It is shown that a classifier achieving 60% recognition on a four

class problem performs as well as an average human labeler. Additionally real life data usually contains multiple emotions at the same time even the conflictual ones [3].

1.1. Previous Research

Previous studies on emotion recognition are done using different data sets with different classifiers and for different number of emotion categories. The research on emotion recognition can be roughly seen as going from acted, speaker dependent, prototypical data to more spontaneous, more realistic and speaker independent data.

Some studies are performed on acted data. In [4], the database investigated consists of studio recordings of a dubbing actress. Three different classifiers are built: *i*) spectral features and GMM, *ii*) prosodic features and SVM, *iii*) prosodic features and GMM. It is observed that the classifier with spectral features outperformed the ones that use only prosodic information. However, SVM classifier with only 6 features achieved an accuracy slightly worse than the best classifier, and this can be preferable because of its computational advantage. In [5], a Spanish database is recorded in studio environment by a professional male actor. Acoustic features and time-frequency representation features are used to build a Bayesian classifier which resulted in 94.6% accuracy. In [6], speaker independent recognition of emotions is aimed. For this purpose 4 isolated HMM classifiers are implemented with different subsets of features. Applying an improved ranked voting method for combining individual HMM classifiers resulted in better accuracies than a single classifier trained with all features.

In [7], the dataset is comprised of acted speech of short utterances. Neural networks, SVM, K-Nearest Neighbors, and Decision trees are implemented using utterance level features. Over 90% accuracy is achieved for distinguishing anger from neutral utterances. In [8], acted data is investigated in terms of 4 emotion classes. All features are based on pitch contours. The pitch contours are smoothed using a spline approximation. Population hillclimbing, promising first selection and forward selection is applied in search of better performing feature sets. Using the selected feature sets Maximum Likelihood Bayes classifier, Kernel Regression and K-nearest neighbor classifiers are im-

plemented. In [9], phoneme-level modeling is proposed for modeling emotions. When utterances of a semi-professional actress is investigated it is observed that the formants of vowels were changing in specific directions for specific emotions. Implementing a phoneme-class dependent HMM resulted in 75.6% recognition accuracy.

In [10], an emotional Turkish dataset is created by 11 different speakers uttering 11 different sentences in four different emotions. Based on pitch, energy, MFCC and zero crossing rate 17 features are extracted. Using a SVM classifier resulted in 70% accuracy, with a false alarm ratio less than 30%.

In [11], both acted emotions and real emotions simulated with a Wizard of Oz (WOZ) setup are investigated. In a WOZ setup subjects interacting with a multimodal dialogue system are observed without knowing that their emotional state was observed. Among the derived 1000+ features from utterances, most related ones are selected by removing correlated features. It is found that best performing feature set for acted speech was predominantly pitch related whereas for spontaneous emotions it is MFCC related.

In recent years many studies have focused on interactive voice response system data in order to enhance human-machine interaction. In [12], data is obtained from a commercial call center where real users interact with a machine agent over telephone. Acoustic, language and discourse information are utilized for detecting negative and non-negative emotions. In order to extract linguistic characteristics emotional salience is introduced which measures the correlation of each word to an emotion category. Additionally, to utilize discourse information responses are categorized into five groups: rejection, repeat, rephrase, ask-start over and none of the above. In [13], German voice portal data is investigated in order to classify anger vs non-anger. It is stated that there exists a considerable amount of garbage turns and modeling of these garbage turns can enhance the classifier accuracy. Also when SVM and GMM classifiers are compared, the former gave better results.

In [14], an emotion detection system on voicemails is created. Four different axes

of emotions as valence (happy vs. sad), arousal (excited vs. calm), urgency, formality are considered. Features extracted from first ten seconds of voicemails are utilized to implement HMM-based classifier. The classifier performed better than chance level for all four axes.

Also human-human dialogs are investigated by Vidrascu and Devillers. In [3], 82% correct classification between negative and positive emotions is achieved on French medical call center data using paralinguistic features. Two labels as major and minor per segment is used during annotation phase. When these labels are analyzed it is found that some of the utterances contained mixture of emotions. In [15], 56% detection is achieved for five emotion classes using acoustic features as well as information extracted from orthographic transcriptions of utterances including disfluencies, affect bursts and phonemic alignment.

In [16], both human-human dialogs recorded during meetings and human-machine data collected through a voice portal is investigated. Three different GMMs are trained using frame level features. The classifiers employed MFCCs, MFCC-low (calculated between 20 and 300 Hz) and pitch features. It is observed that the two MFCC classifiers achieved similar accuracies while MFCC-low outperformed the pitch features. The system performance improved significantly when the scores of the three classifiers are combined using multiple linear regression.

Emotion recognition research is generally done using different data sets with different experimental setups which makes results incomparable. Also most databases are far from being realistic by containing acted data recorded in studio conditions. To provide a more spontaneous and less prototypical data set with clearly defined train and test sets Fau Aibo emotion corpus [17] is created and made publicly available. For baseline experiments, low level descriptors (LLD) such as pitch, energy and mel frequency cepstral coefficients (MFCC) are utilized. Implementing a Hidden Markov Model (HMM) classifier on the frame level resulted in 66.1% and 35.9% unweighted average recall (UA) values for 2-class and 5-class problems respectively. SVM with same LLD on the chunck level resulted in 67.7% and 38.0% UA recall for 2-class and

5-class problems respectively.

Fau Aibo emotion corpus is studied in several papers. In [18], acoustic and linguistic features are fused in order to detect angry vs non-angry utterances. Feature selection by information gain ratio criterion showed that spectral features and power-related features are most valuable to classification. SVM is used for acoustic classification while emotional saliance is employed for linguistic classification. When the two decisions are fused an UA recall of 67.6% is achieved. In [19], GMM classifier is built using prosodic, spectral as well as HMM-based features. Fusing different scores obtained for different feature sets 67.9% and 41.6% UA is achieved for 2-class and 5-class cases respectively.

In [20], two linguistic, two acoustic and an acoustic-linguistic classifier are implemented using Fau Aibo dataset considering the 2-class problem. Two different types of late fusion techniques are employed. Democratic majority voting resulted in 70.35% UA accuracy. Using the additional confidence scores of different classifiers, Kononenko discretization with added features are applied for fusion by learning. An UA of 70.45% is achieved which is only slightly better than majority voting.

In order to investigate the effect of real life conditions on emotion recognition Berlin Emotional Speech Database, Danish Emotional Speech Database, and AIBO Emotion Corpus are used in [21]. Obtaining a SNR value of 0 dB by addition of white noise degraded the performances approximately 10% for both acted and spontaneous databases. For investigating the influence of microphone condition and room acoustics experiments are implemented by close talk, close talk reverberated, and room microphone recordings of Aibo dataset. Highest accuracies are achieved for the first case while the lowest performances are achieved for the last setup. Additionally it is observed that recognition of speaker independent spontaneous emotions are more difficult than studio quality acted emotions.

Prosodic parameters are proven to be good correlates for emotions as well as being applicable to all kinds of data. However, linguistic parameters are only applicable

to spontaneous data since content of acted data is predetermined. In order to capture lexical information, language models are trained [22], emotional salience of words are calculated [12, 18, 23] and classification with a bag of words approach [23] is implemented. Additionally belief networks are trained for their ability to handle uncertain information supplied by automatic speech recognizer [24, 25].

In [26], different linguistic classifiers are compared using data extracted from a deployed customer-care system, the AT&T "How May I Help You" system (HMIHY 0300). Three classifiers are built considering negative and non-negative emotions, using the output of an automatic speech recognition system which has a word error rate of 37.8%. The interpolated language model classifier, the mutual information-based classifier and kernel based classifiers achieved 70.1%, 78.8% and 80.6% classification accuracies respectively.

Since Turkish is an agglutinative language, words are generated by appending suffixes to roots. This morphological structure makes it difficult to build robust word-based language models. Also reasonable size dictionaries suffer from high out of vocabulary (OOV) words. The solution is using sub-word units instead of words. This type of language models are applied for automatic speech recognition applications [27]. Additionally stemming (removing suffixes from roots) is found to be successful for document retrieval [28]. Therefore it will be interesting to investigate the different language models in terms of emotion recognition.

1.2. Outline of the Thesis

In this thesis we investigated a Turkish call center dataset and German Fau Aibo dataset. Binary emotion classification is considered. SVM and GMM classifiers are built using acoustic features. Linguistic parameters are extracted from manual transcriptions of recordings. To observe the effect of sub-word models and stemming in emotion recognition task three different models are built for calculating language model scores; *i*) word based *ii*) stem-only *iii*) stem+ending. Then fusion at the decision level is implemented using a multi layer perceptron (MLP).

This thesis is organized as follows: In Chapter 2, theoretical background of the classifiers is presented. The databases investigated are presented in Chapter 3. Data analysis considering acoustic, spectral and linguistic characteristics are given in Chapter 4. In Chapter 5, proposed system, features used and classifier implementations are explained. In Chapter 6, results for different classifiers with their comparison and discussion of the obtained results are given. The conclusion of the thesis is in the final chapter.

2. THEORETICAL BACKGROUND

2.1. Feature Extraction

The block diagram of a basic emotion recognition system is depicted in Figure 2.1. Emotion recognition systems are generally based on two main stages. First features are extracted from speech signals and then a classifier is designed using these features.

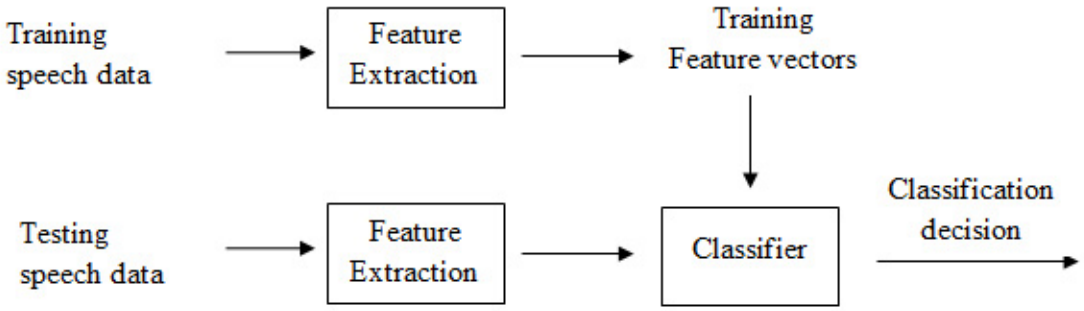

Figure 2.1. Block diagram of an emotion recognition system.

When emotionally colored speech is investigated it is observed that pitch, intensity, duration, and spectral properties were changing with emotional states of the speakers [29]. Therefore different statistical properties of mentioned characteristics of utterances are utilized as features.

For modeling spectral characteristics Mel Frequency Cepstral Coefficients are frequently adopted for speech processing applications including emotion recognition, speech recognition, and voice verification. MFCCs cover the spectrum with nonlinear filter banks which is similar to the human auditory system. The mapping of the real frequencies to mel scale is approximated by

$$mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.1}$$

To calculate the MFCCs, Fourier Transform is applied to windowed frames of the signal. Then the powers are obtained for overlapping triangular regions. These frequency

bands are shown in Figure 2.2 which is formed by mel scale.

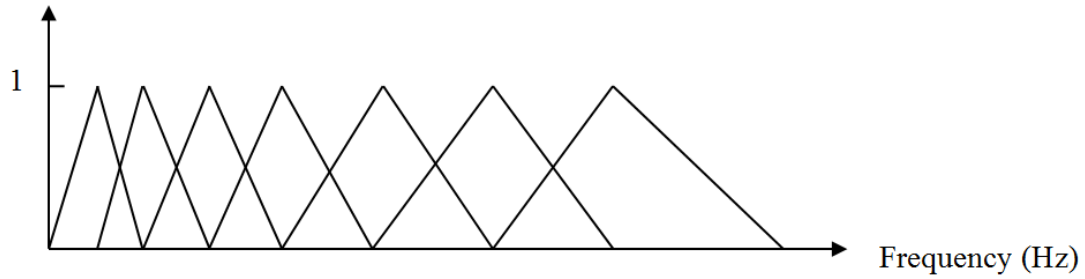

Figure 2.2. Mel-Scale Filter Banks.

Applying discrete cosine transform to powers in frequency bands and taking the amplitudes results in MFCCs. The discrete cosine transform is formulated as

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j cos(\frac{\pi i}{N}(j - 0.5)) \tag{2.2}$$

where m_j are the log filter bank amplitudes and N is the number of frequency bands.

2.2. Modeling

2.2.1. Support Vector Machines

Support Vector Machines are discriminative classifiers which aim to find separating hyperplane with maximum margin between two classes. A sample SVM setup is given in Figure 2.3.

Here y_i is 1 or -1 which indicates the class of vector $\vec{x_i}$. The parallel hyperplanes that maximize the margin while separating the two classes can be written as

$$\vec{w} \cdot \vec{x} - b = 1 \tag{2.3}$$

$$\vec{w} \cdot \vec{x} - b = -1 \tag{2.4}$$

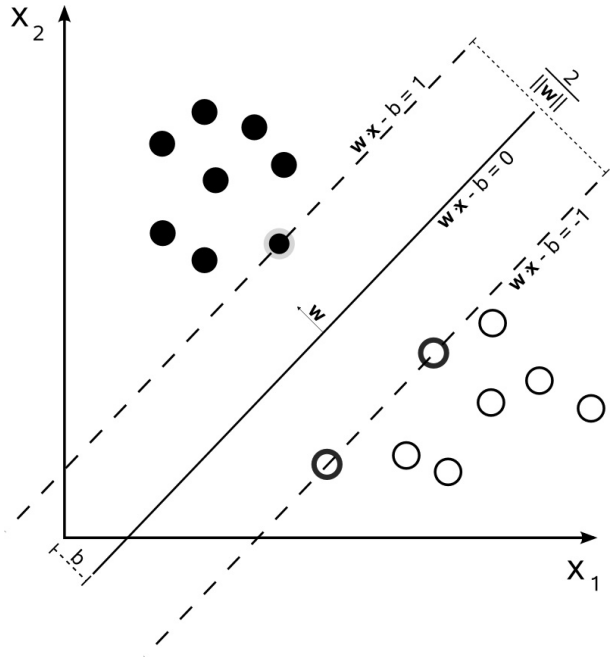

Figure 2.3. A sample SVM setup.

Assuming linear separability, the difference between the hyperplanes is found to be $2/\|\vec{w}\|$. In order to maximize the distance, $\|\vec{w}\|$ should be minimum. Also

$$\vec{w} \cdot \vec{x} - b \geq 1; \text{ for } \vec{x_i} \text{ of the first class} \tag{2.5}$$

$$\vec{w} \cdot \vec{x} - b \leq -1; \text{ for } \vec{x_i} \text{ of the second class} \tag{2.6}$$

which can be rewritten as

$$y_i(\vec{w} \cdot \vec{x} - b) \geq 1 \tag{2.7}$$

Using Lagrange multipliers this constrained problem can be expressed as

$$\min_{(\vec{w},b)} \max_{\alpha} \{\frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\vec{w} \cdot \vec{x_i} - b)]\} \tag{2.8}$$

Solving this by quadratic programming techniques results in

$$\vec{w} = \sum_{i=1}^{N} \alpha_i y_i \vec{x_i} \tag{2.9}$$

When samples are not linearly separable slack variables, ξ_i, are introduced. The aim is to select the hyperplane which separates the samples as much as possible while the margin is maximized. Then the optimization problem becomes:

$$\min_{\vec{w},\xi} \{\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N} \xi_i\} \tag{2.10}$$

subject to

$$y_i(\vec{w} \cdot \vec{x} - b) \geq 1 - \xi_i \tag{2.11}$$

In order to model nonlinearities original vectors are mapped to a different space by kernel functions. Some common kernels are:

- Polynomial (homogeneous): $K(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j})^d$
- Polynomial (inhomogeneous): $K(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + 1)^d$
- Gaussian Radial Basis Function: $K(\vec{x_i}, \vec{x_j}) = \exp(-\gamma\|\vec{x_i} - \vec{x_j}\|^2)$, for $\gamma > 0$.
- Hyperbolic tangent: $K(\vec{x_i}, \vec{x_j}) = \tanh(k\vec{x_i} \cdot \vec{x_j} + c)$, for some $k > 0$ and $c < 0$.

2.2.2. Gaussian Mixture Model

Gaussian Mixture Models are representation of probability distributions as weighted average of several Gaussians. The likelihood contributed by i'th Gaussian component for a d dimensional vector $\vec{x}$ is

$$p(\vec{x}|\vec{\mu}_i, \mathbf{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} * \exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)'\mathbf{\Sigma}_i{}^{-1}(\vec{x} - \vec{\mu}_i)\} \tag{2.12}$$

where $\vec{\mu}_i$ is mean and $\mathbf{\Sigma}_i$ is covariance matrix of i'th Gaussian. The total likelihood for the given model is calculated as

$$p(\vec{x}) = \sum_{i=1}^{N} w_i p(\vec{x}|\vec{\mu}_i, \mathbf{\Sigma}_i) \tag{2.13}$$

where N is the number of Gaussians and w_i is the weight of Gaussian i with the constraint

$$\sum_{i=1}^{N} w_i = 1 \text{ and } \forall i : w_i \geq 0. \tag{2.14}$$

GMM is trained by Expectation-Maximization algorithm which iteratively increases the likelihoods until a predetermined convergence criterion is reached.

2.2.3. Language Modeling

Language modeling is calculating the probabilities of sequences of words. For this purpose generally the probabilities of word sequences $w_1, ..., w_m$ are approximated by n-grams as,

$$P(w_1, ..., w_m) = \prod_{i=1}^{m} P(w_i|w_1, ..., w_{i-1}) \approx \prod_{i=1}^{m} P(w_i|w_{i-(n-1)}, ..., w_{i-1}) \tag{2.15}$$

where it is assumed that observing a word, w_i, depends only on the preceding n-1 words. The conditional probabilities are calculated by frequency counts

$$P(w_i|w_{i-(n-1)}, ..., w_{i-1}) = \frac{count(w_{i-(n-1)}, ..., w_i)}{count(w_{i-(n-1)}, ..., w_{i-1})} \tag{2.16}$$

Frequently used n-grams are unigrams, bigrams and trigrams which imply n=1, n=2, and n=3 respectively. However only using frequencies results in assigning zero probabilities to all unseen sequences of words which are also probable in real life. Therefore smoothing techniques are applied in order to generalize the probability distributions calculated by n-grams. A simple smoothing technique is “add one smoothing” which

approximates the conditional probabilities as:

$$P(w_i|w_{i-1}) \approx \frac{count(w_{i-1}w_i)+1}{count(w_{i-1})+V} \tag{2.17}$$

where V is the total number of observed words. A modified version is "add delta smoothing",

$$P(w_i|w_{i-1}) \approx \frac{count(w_{i-1}w_i)+\delta}{count(w_{i-1})+V\delta} \tag{2.18}$$

Good-Turing smoothing is based on the idea that number of singletons is correlated to number of novel events.

$$P(w_i) \approx \frac{r^*}{N} \tag{2.19}$$

$$r^* = (count(w_i)+1)\frac{n_{r+1}}{n_r} \tag{2.20}$$

where n_r is the number of words that are observed r times and n_{r+1} is the number of words that are observed $r+1$ times. There are other smoothing techniques such as Katz, Jelinek-Mercer, Absolute Discounting and Kneser-Ney smoothing which are applicable to higher order n-grams than unigrams.

2.2.4. Artificial Neural Networks

A neural network (NN) is a modeling tool that can represent complex relationships between input and output. The computation is distributed among artificial neurons. Artificial neurons have multiple inputs from other neurons which are weighted and summed before a transfer function is applied. Acquiring information by learning and storing information through weights between the neurons resembles to the mechanism of brain. The training can be done in a supervised manner, unsupervised manner or by reinforcement learning.

Commonly used neural networks are multi layer perceptrons (MLP) which are

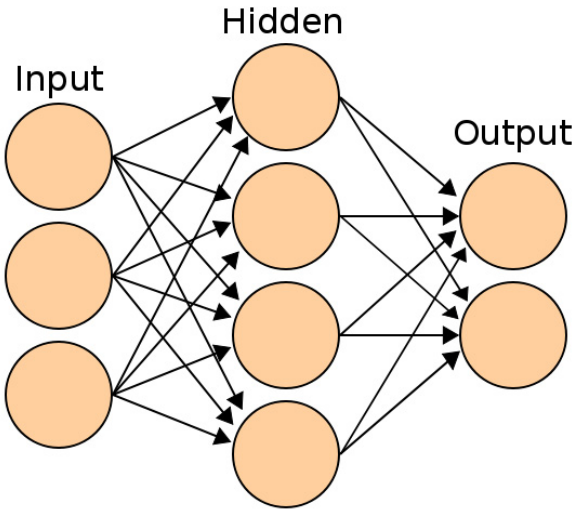

Figure 2.4. A sample ANN setup.

feedforward NN's. A MLP creates a mapping from given input vector to the desired output. It consists of nodes with nonlinear transfer functions. Outputs of nodes in a layer are weighted and added while being delivered to the next layer. Backpropagation algorithm is used to train MLP which minimizes the cost by gradient descent.

2.2.5. Evaluation Metrics

For a two class problem classification results can be categorized as in Table 2.1.

Table 2.1. Classification results for a two class problem.

	Correct Classification	
	Class1	Class2
Obtained Class1	true positive (tp)	false positive (fp)
Obtained Class2	false negative (fn)	true negative (tn)

Using these categories precision, recall and accuracy are defined as:

$$\text{Precision} = \frac{tp}{tp + fp} \tag{2.21}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{2.22}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.23}$$

For an unequally distributed pattern recognition application a classifier deciding always in favor of the majority class can result in the highest accuracies. As a result accuracy and weighted average recall values do not reflect the true performance of the system. Therefore unweighted average recall values are frequently employed.

3. DATABASES

3.1. Call Center Dataset (Turkish)

The first database used in this thesis is Call Center Dataset (CCD). It consists of utterances of real agent-client dialogs recorded in 3 Turkish commercial call centers from finance, insurance and telecommunications sectors.

The dataset to be labeled is filtered out of 300 hours of conversations. Agents have indicated possibly problematic calls. This remaining set comprised of 6 hours 13 minutes of 8 kHz, 8 bit mu-law encoded audio. Using a voice activity detection (VAD) module, these 385 dialogs are split into 8512 utterances with durations ranging between 0.5 and 31 seconds. Then the utterances are separated into 4 non overlapping subgroups. Each subgroup is labeled by a different person. The labelers marked each utterance as angry, non-angry or garbage. Garbage label is assigned for turns such as silences, DTMF tones and overlapping speech. These data are not included in experiments.

Data are separated into train and test sets as in Table 3.1. There is no overlap in terms of speakers between train and test sets.

Table 3.1. Number and duration of utterances for train and test sets.

	angry		non-angry	
	# of utterances	total duration	# of utterances	total duration
Train set	1052	01:08:40	6531	04:28:06
Test set	146	00:08:10	783	00:23:08
$\sum$	1198	01:16:50	7314	04:51:14

Apart from 4 labelers 2 other labelers labeled the test set for measuring labeler-

labeler agreement. Kappa statistic is used for this purpose as in [12, 13].

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{3.1}$$

where $P(A)$ is the probability of agreement between labelers and $P(E)$ is the probability of agreement by chance. The kappa value being less or equal to 0 means no agreement and 1 means total agreement between the labelers. Kappa value between the two labelers is found to be 0,79 which indicates a high degree of conformity.

3.2. FAU Aibo Dataset (German)

The second database used in this thesis is FAU Aibo Dataset (FAD) which consists of recordings of children playing with Sony's pet robot Aibo [30, 31]. The corpus is formed of spontaneous German speech recorded at two different schools, Mont and Ohm. The children interacting with the robot thought that Aibo responded to their commands while it was being controlled by a human operator. The robot sometimes behaved disobediently to the commands for provoking emotions.

There is a total of 9.2 hours of speech data without pauses collected from 21 male and 30 female students aged 10-13. The speech is recorded by wireless headsets in 48 kHz 16 bit format and then resampled to 16 kHz. After resampling the utterances were split into turns with a pause threshold of 1 seconds.

Table 3.2. Number of utterances for train and test sets.

	Negative	Idle	$\sum$
Train set	3,358	6,601	9,959
Test set	2,465	5,792	8,257
$\sum$	5,823	12,393	18,216

Five labelers annotated the data at the word level. From 11 fine grained emotion labels the two-class problem is formed as Negative and Idle. The negative class includes angry, touchy, reprimanding, and emphatic utterances while the remaining utterances

constitute the idle class. Final decision for the word is made by majority voting. Chunk level labels are constructed using raw labels of all five annotators. A chunk is considered negative if at least 50% of the raw labels are negative or it satisfies following two conditions: *i*) at least one third of raw labels are negative. *ii*) at least 90% of remaining labels are either negative or neutral.

Experiments on a subset of data revealed that chunks are better analysis units then words and turns on this corpus [30]. The distribution of data for negative and idle emotions for train and test sets is given in Table 3.2. In order to achieve speaker independence the data of school Ohm is used for training while the data of school Mont is used for testing.

4. DATA ANALYSIS

4.1. Prosodic Characteristics of Real Emotions

Pitch and energy contours of three angry and non-angry utterances are given in Figures 4.1 and 4.2.

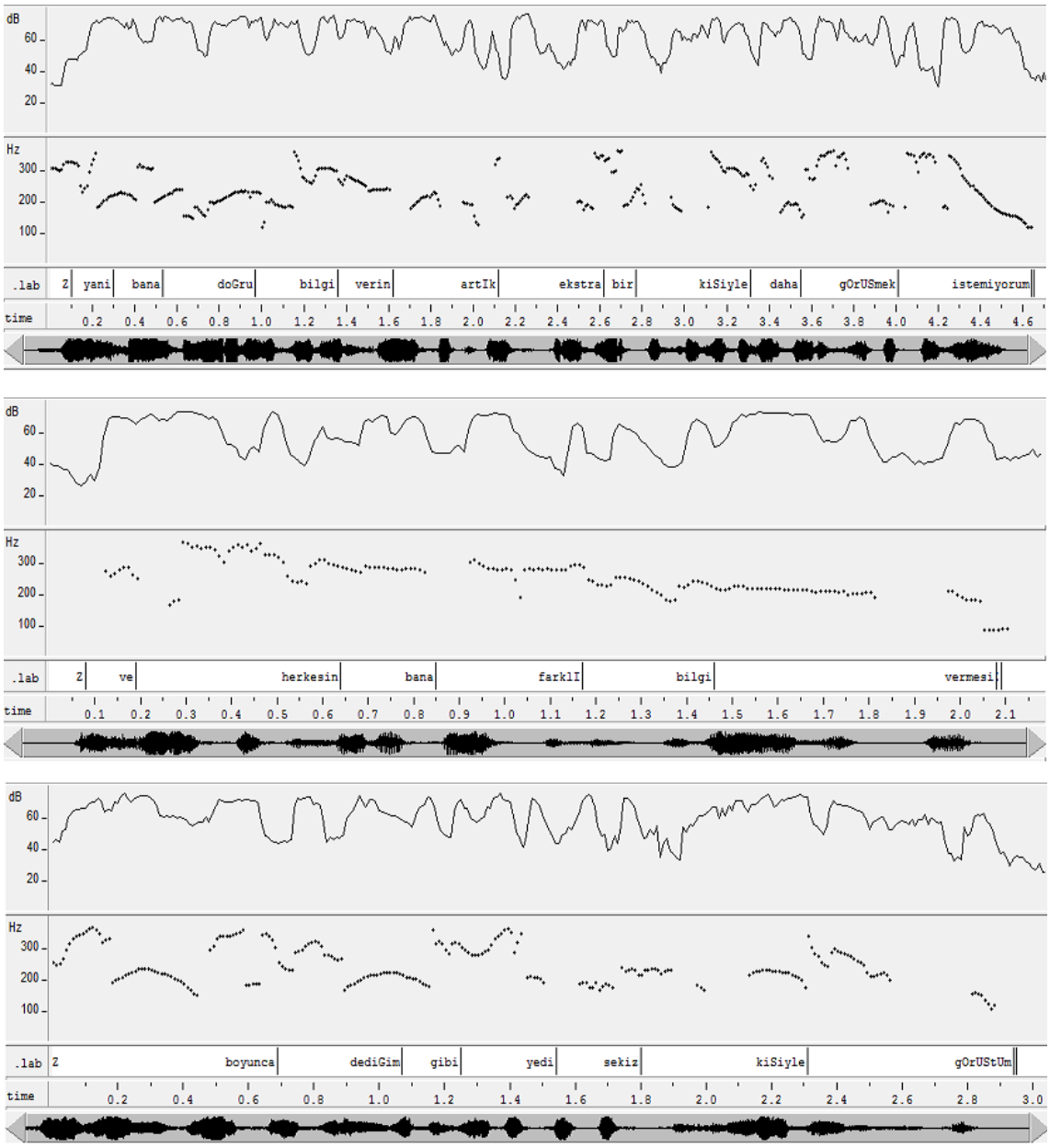

Figure 4.1. Three angry utterances with pitch and energy contours.

When pitch and energy contours for angry and non-angry utterances are exam-

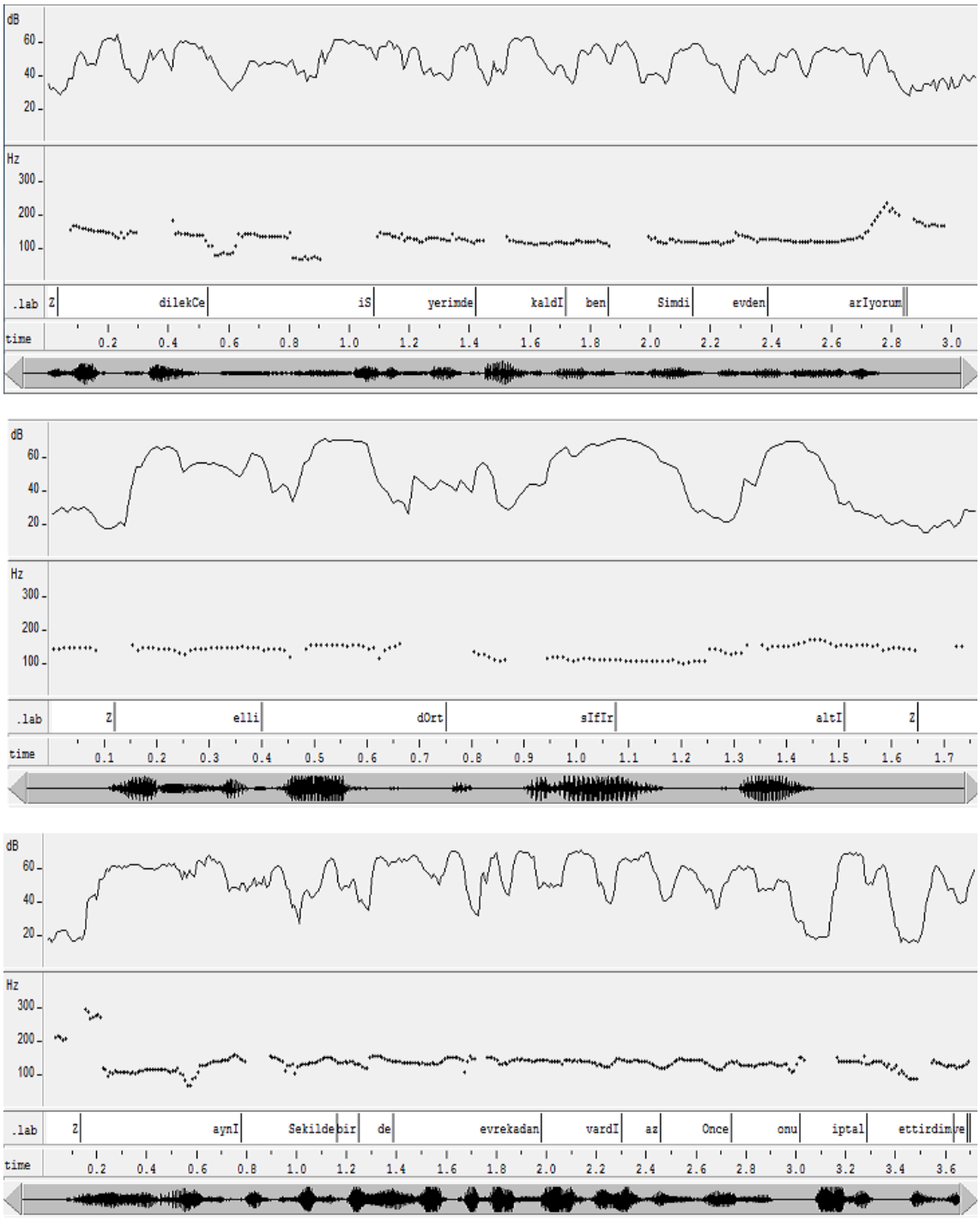

Figure 4.2. Three non-angry utterances with pitch and energy contours.

ined we observe that range, variations and variations of slopes are higher for angry utterances. The maximum difference within a voiced segment (intra-voiced) as well as between consecutive voiced segments (inter-voiced) for the pitch contours of angry utterances are greater than that of normal utterances. Also there is a decay of pitch at the endings of angry sentences whereas non-angry sentences end with flat pitch

contours. Additionally the pitch contours of angry utterances contain arc like shapes which will reflect as sign changes in the slopes of the pitch contours.

4.2. Prosodic Characteristics of Simulated Emotions

Pitch and energy contours of three sentences uttered by a person with angry and non-angry emotions are given in Figures 4.3 and 4.4.

Same sentences are recorded from the same person in order to allow direct comparison. Similar to the real emotion characteristics, the range, variations and variations of slopes are higher for angry utterances.

4.3. Spectral Characteristics

Spectograms of utterances for words “aibo” and “stopp” with negative and idle emotions are presented in Figures 4.5 and 4.6. When we compare the spectograms we observe that higher frequency components are boosted for negative emotions.

4.4. Linguistic Characteristics

Relative frequencies of all words as well as their differences for the two datasets used are presented in Figures 4.7 and 4.8. Since the context is limited to a set of commands on FAD, the vocabulary size is relatively small. As a consequence the average differences of relative frequencies for FAD is smaller than CCD average.

Words more related to anger class on CCD are listed in Table 4.1 considering the difference between the relative frequencies of the two emotions. High correlation of conjunction words “ve” (and), “ki” (that) and “yani” (namely) to anger class can be attributed to subjects effort to construct longer sentences and inhibit the interruption of the agent. Also the word “iptal” (cancel) may indicate the resign of the costumer from a service or all the products that the company offers which explains the high correlation to anger class. Another inference can be made from words “ben” (I), “benim” (my),

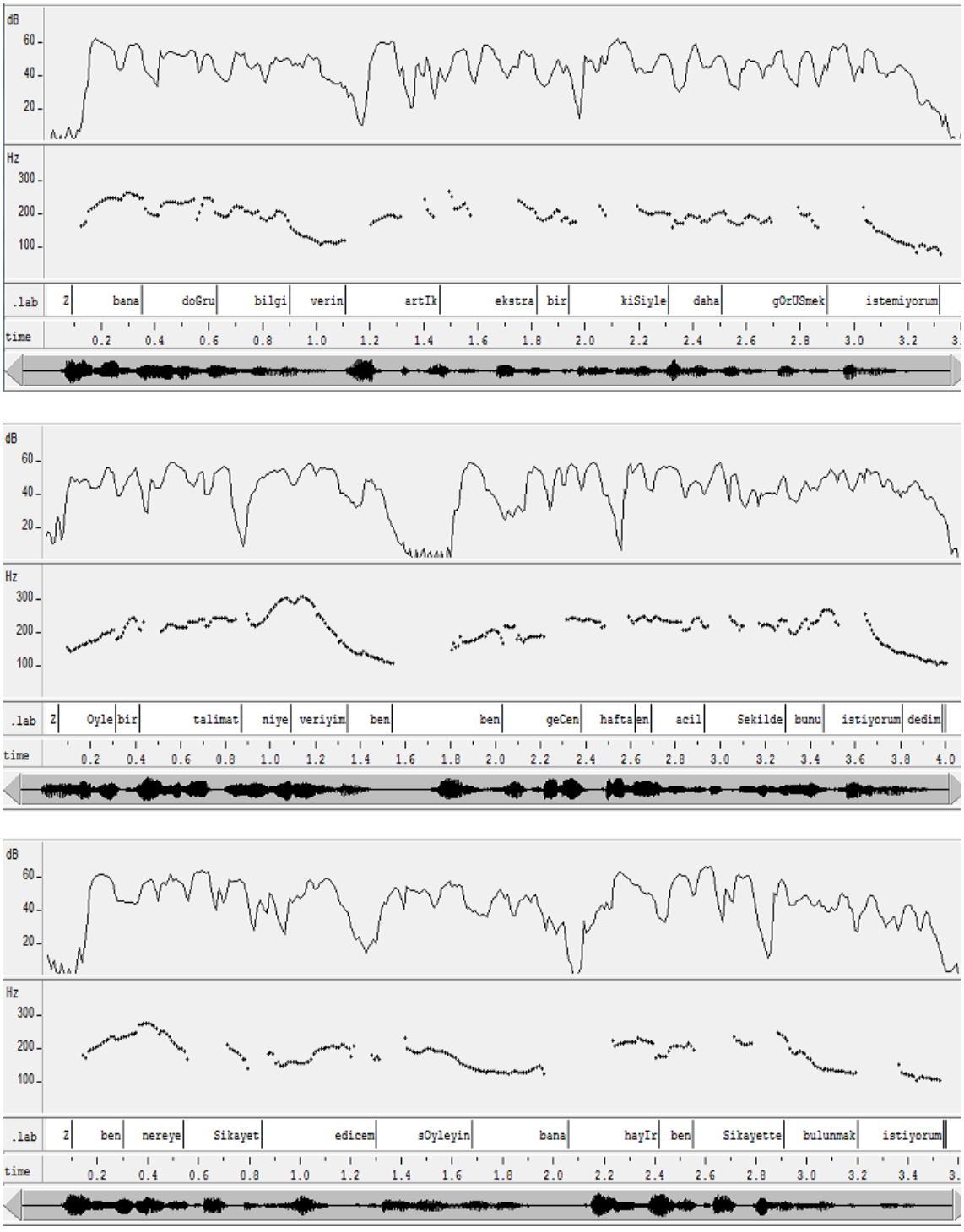

Figure 4.3. Three simulated angry utterances with pitch and energy contours.

"bana" (to me), and "beni" (me). We can conclude that angry sentences are mostly constructed with subjects and objects as the talking person.

Words more related to non-anger class on CCD are listed in Table 4.2. Words

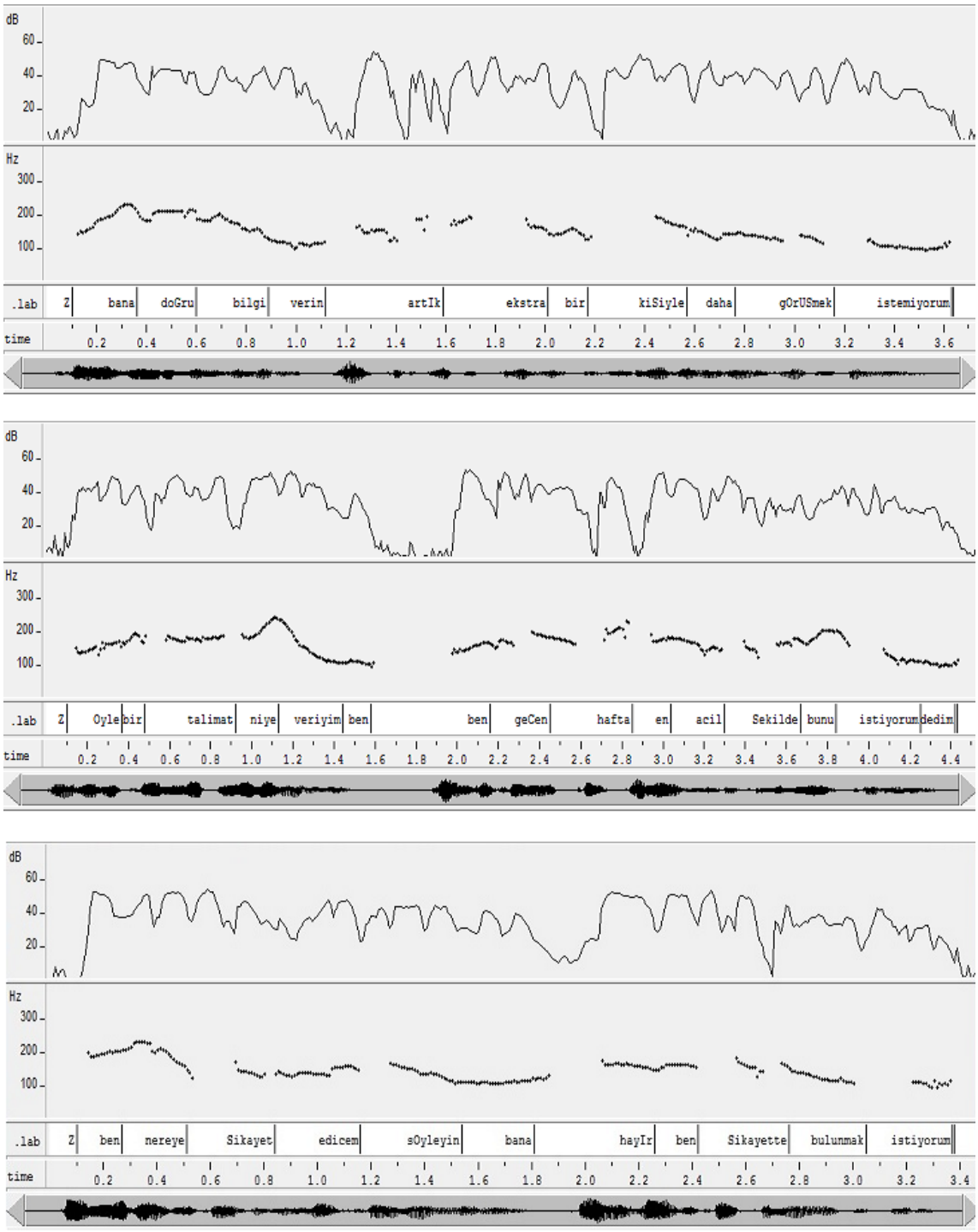

Figure 4.4. Three simulated non-angry utterances with pitch and energy contours.

"evet" (yes), "iyi" (good), "teşekkür" (thanks) and "tamam" (okay) reflects the content and approval of the subjects. Also numbers are mostly related to personal information exchange which makes utterances containing them to be more related to neutral emotional state.

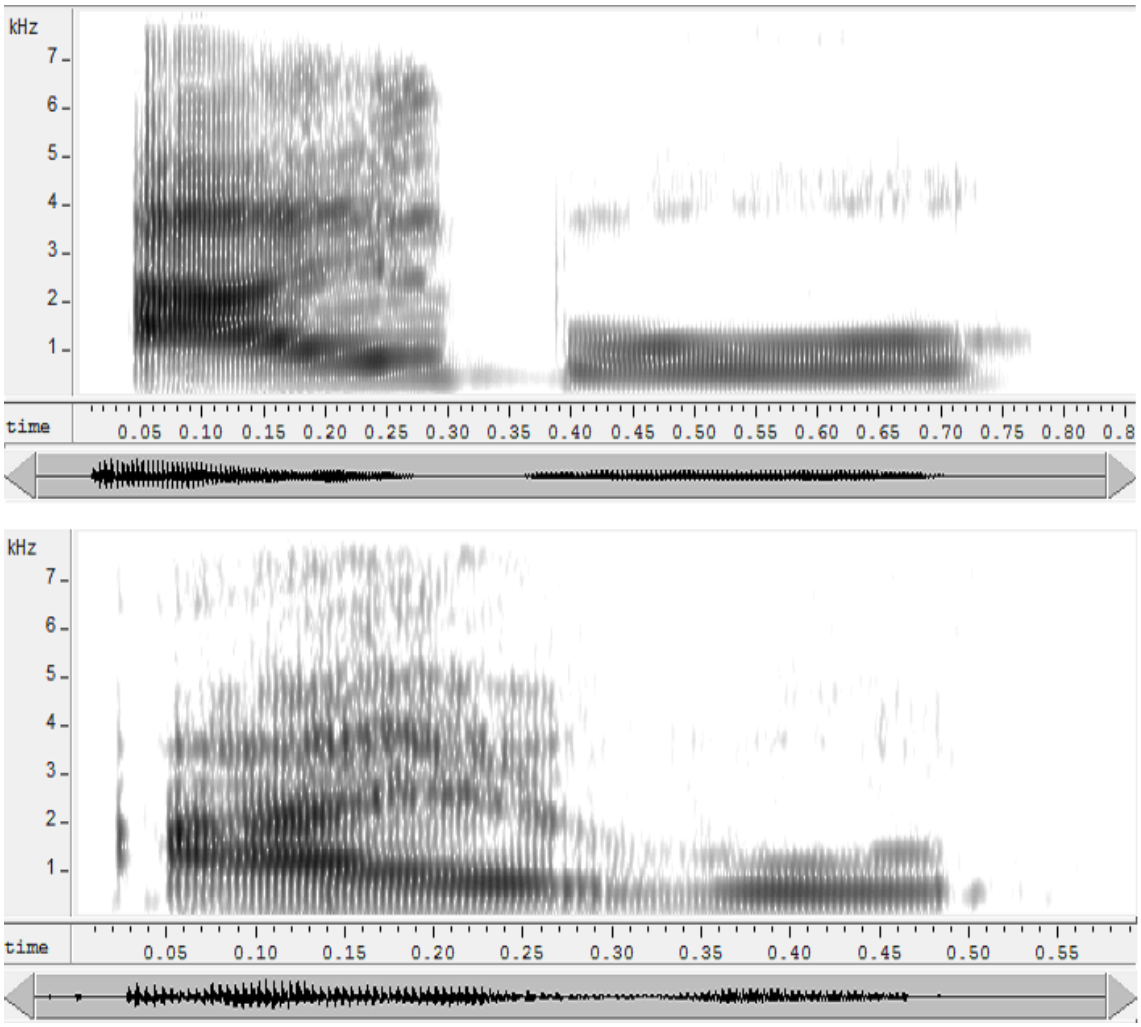

Figure 4.5. The word “aibo” negative (top) and idle (bottom).

Words more related to negative class on FAD are listed in Table 4.3. Words indicating dissatisfaction such as “stopp” (stop) and “nein” (no) implies the subjects’ disapproval of the action which may result in negative attitudes.

Words more related to idle class on FAD are listed in Table 4.4. Words such as “ja” (yes), “gut” (good), “weiter” (more) and “okay” (okay) reflect the approval and satisfaction of the subject.

Endings more related to anger and non-anger classes on CCD are listed in Tables 4.5 and 4.6. We observe that ending “_ın” which indicates impoliteness, is related

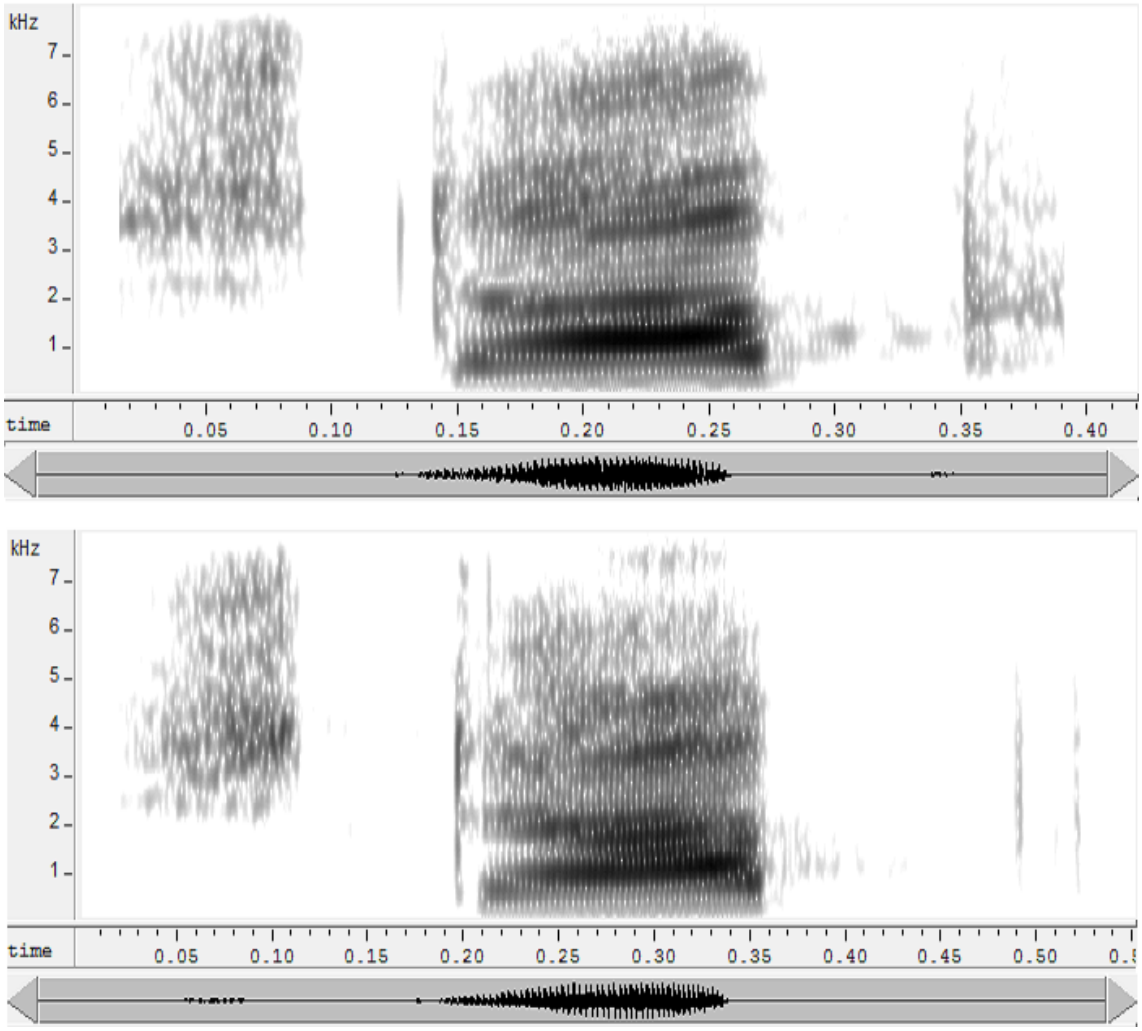

Figure 4.6. The word “stopp” negative (top) and idle (bottom).

to anger class.

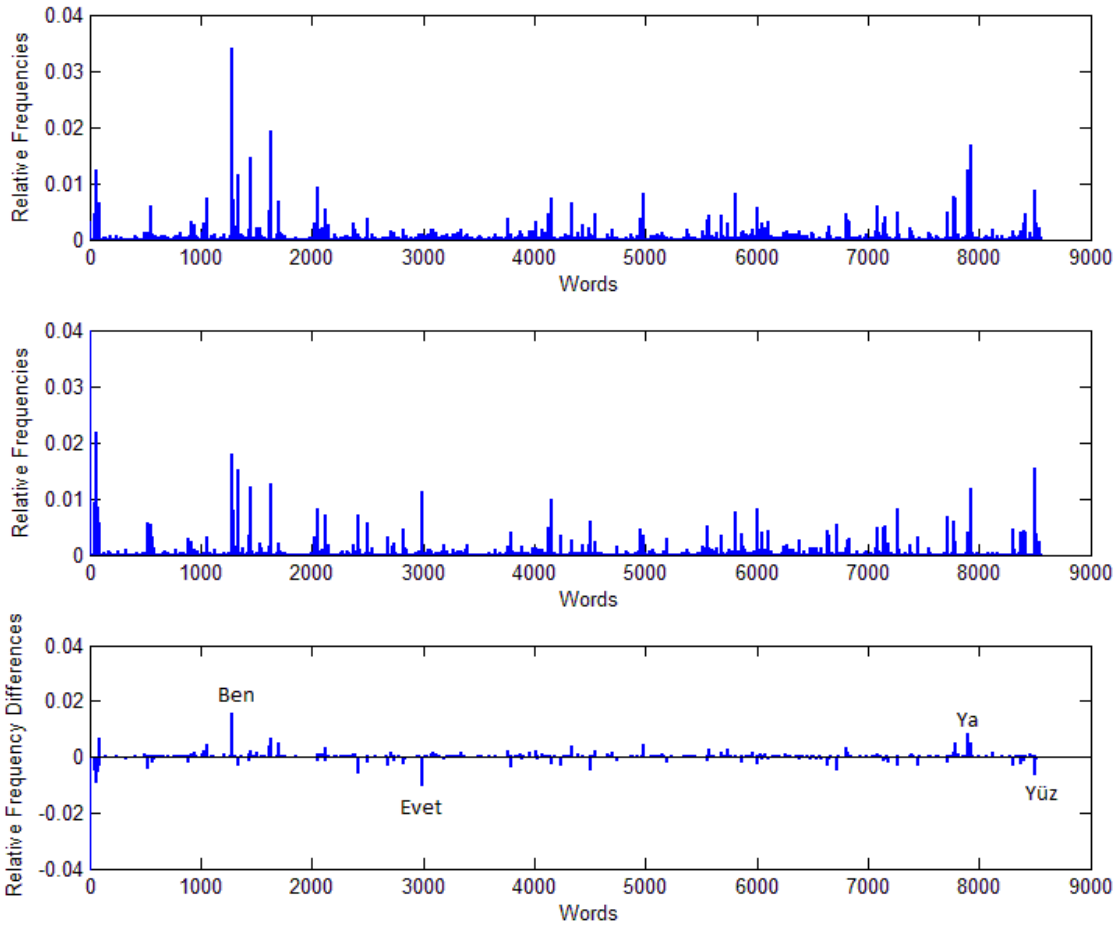

Figure 4.7. Relative frequencies of all words for CCD. Anger (top), Non-Anger (middle), Difference between anger and non-anger (bottom)

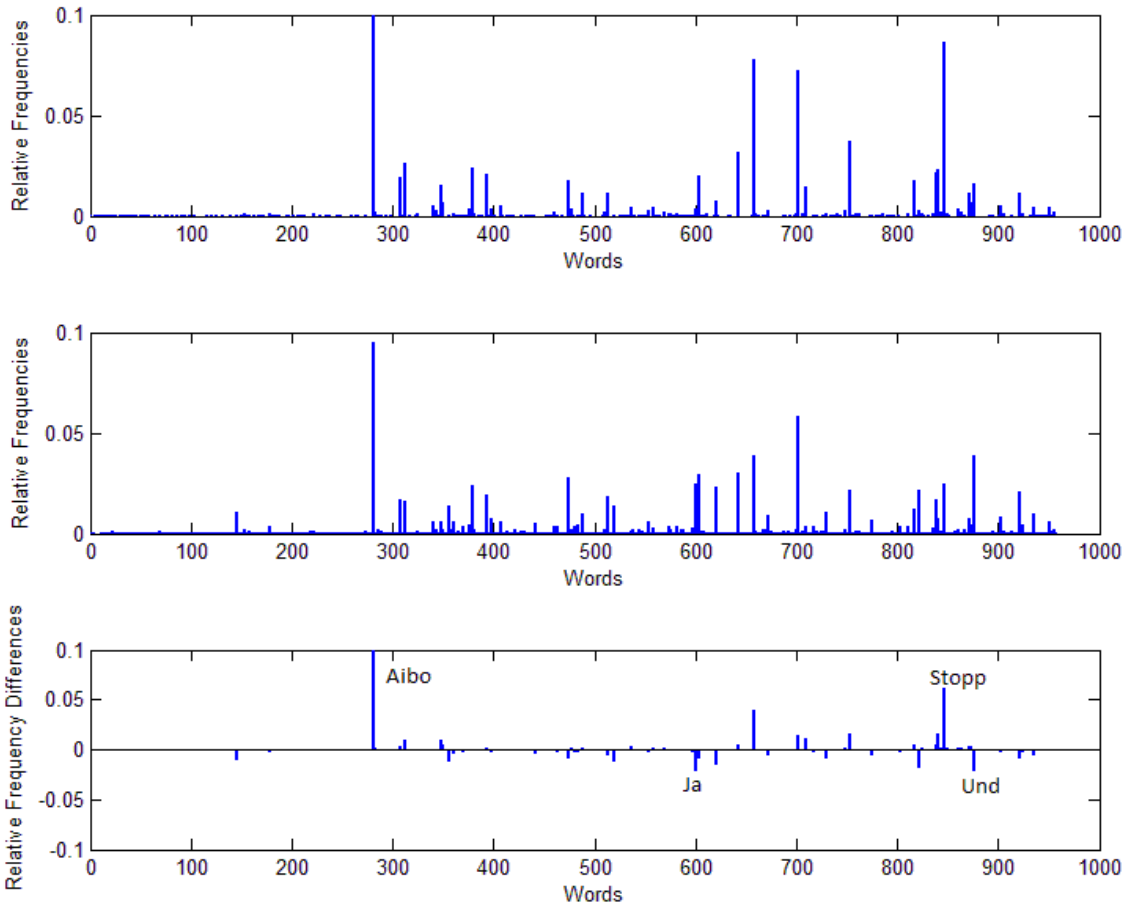

Figure 4.8. Relative frequencies of all words for FAD. Negative (top), Idle (middle), Difference between negative and idle (bottom)

Table 4.1. Words more related to anger class on CCD.

Words	English Meanings	Differences in Relative Frequencies
ben	I	0.01590
ya	well	0.00832
bu	this	0.00680
benim	my	0.00519
ve	and	0.00492
yani	namely	0.00483
bunu	it	0.00480
ki	that	0.004730
bana	to me	0.00426
beni	me	0.00378
iptal	cancel	0.00376
böyle	such	0.00368
siz	you	0.00327

Table 4.2. Words more related to non-anger class on CCD.

Words	English Meanings	Differences in Relative Frequencies
evet	yes	-0.01041
yüz	hundred	-0.00680
dokuz	nine	-0.00596
iyi	good	-0.00506
sıfır	zero	-0.00481
altı	six	-0.00428
günler	days	-0.00413
bir	one	-0.00357
teşekkür	thanks	-0,00333
ilgili	related	-0.00328
tamam	okay	-0.00321
sekiz	eight	-0.00316
yedi	seven	-0.00314

Table 4.3. Words more related to negative class on FAD.

Words	English Meanings	Differences in Relative Frequencies
Aibo	Aibo	0.11431
stopp	stop	0.06218
links	next	0.03935
rechts	right	0.01553
stehen	stand	0.01498
nach	towards	0.01369
nein	no	0.01070
aufstehen	stand up	0.00995
bleib	permanent	0.00930
sitz	sit	0.00485
bleiben	remain	0.00478
laufen	run	0.00475
steh	stand	0.00471

Table 4.4. Words more related to idle class on FAD.

Words	English Meanings	Differences in Relative Frequencies
und	and	-0.02239
ja	yes	-0.02152
so	so	-0.01901
komm	come	-0.01511
gut	good	-0.01329
brav	good	-0.01256
is	is	-0.01032
geh	go	-0.00977
weiter	more	-0.00974
jetzt	now	-0.00963
okay	okay	-0.00918
geradeaus	straight	-0.00699
schon	already	-0.00657

Table 4.5. Endings more related to angry class on CCD.

Endings	Differences in Relative Frequencies
_m	0.00512
_dı	0.00313
_na	0.00266
_ye	0.00223
_nı	0.00177
_ın	0.00103
_dan	0.00097
_mı	0.00097
_um	0.00094
_nız	0.00092
_fendi	0.00091
_ları	0.00082
_lık	0.00069

Table 4.6. Endings more related to non-angry class on CCD.

Endings	Differences in Relative Frequencies
_ler	0.00281
_yı	0.00232
_i	0.00208
_e	0.00206
_lı	0.00074
_yla	0.00065
_bı	0.00065
_un	0.00059
_hangi	0.00049
_ten	0.00046
_lar	0.00043
_halde	0.00037
_nda	0.00033

5. PROPOSED SYSTEM

In this study three different classifiers are trained with three different sources of information. The block diagram of the proposed system is given in Figure 5.1.

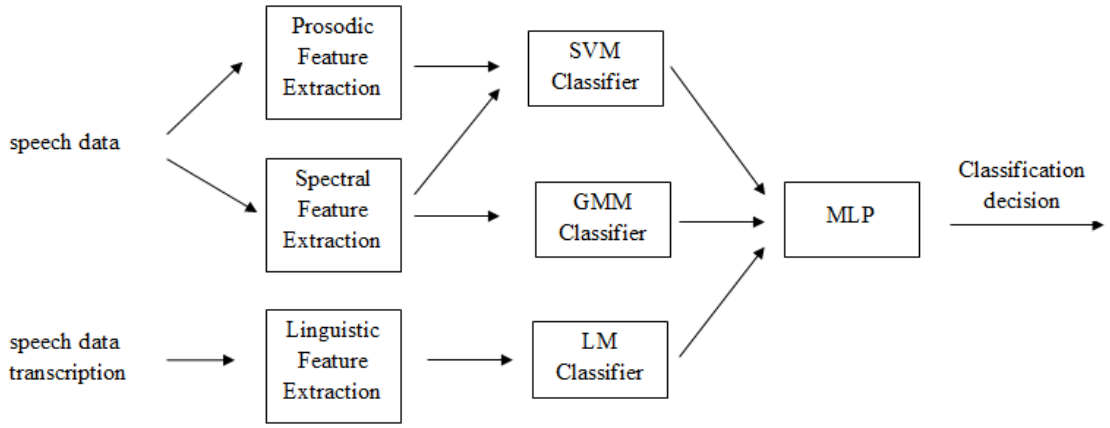

Figure 5.1. Block diagram of the proposed system.

Prosodic and linguistic features are extracted on the utterance level while spectral features are extracted on the frame level. For converting the spectral features to utterance level parameters during SVM classification the averages are calculated over all frames of each utterance.

5.1. SVM

Libsvm library [32] is used for SVM modeling end testing. For SVM experiments 143 features are extracted from each utterance. These are:

- MFCC parameters : 13 coefficients with deltas and delta deltas combines to 39 features. Mean, minimum and maximum values for frames of an utterance results in a 117 dimensional vector.
- Pitch parameters : min, median, max, first quartile, third quartile, mean and max of first derivative, inter voiced maximum difference, intra voiced maximum difference. Pitch contours are extracted by robust algorithm for pitch tracking

(RAPT) [33]. Then z-normalized by shifting the mean to 0 and scaling the variance to 1, for eliminating speaker differences.

- Energy parameters : min, max, mean, standard deviation, mean and max of first derivative.
- Microprosody parameters : min, max, standard deviation of jitter and shimmer. Jitter and shimmer are calculated using a linear filter as used in [7].

During pitch contour calculations it is observed that for noisy recordings background speech contours were also included. To eliminate these errors a bias towards unvoiced decision is introduced.

5.2. GMM

Gaussian Mixture Models for angry and non-angry model classes are trained with spectral features. Spectral features are extracted using HTK [34] toolkit. 13 Mel Frequency Cepstrum Coefficients are calculated for 25 ms Hamming windowed frames with 10 ms skip rate. With the addition of delta features a 26-dimensional vector for each frame is formed. We used 64 mixtures for each GMM.

Becars software [35] is used for GMM training end testing.

5.3. LM

For language modeling all utterances are manually transcribed. Only words and human noises are included in transcriptions. In order to split words into sub-word parts Morfessor[36] is used. Morfessor is an unsupervised morphological analyzer that segments words into *morphs* which are morpheme-like units. First morph of a word is labeled as a stem while remaining morphs are concatenated and labeled as endings. Also endings are marked by a special character to avoid possible confusions. After creating stemmed and stem+ending forms, the language model processing applied is the same as word model processing.

Table 5.1. Language model train set lexical analysis for CCD.

	Angry	Non-angry
# of words	10259	38390
# of distinct words	2985	6953
# of distinct stems	1922	3913
# of distinct morphs	2387	4788

For language modeling unigrams are employed since it is assumed that emotions mostly affect the choice of words and order of words is not relevant. Two different unigrams are trained for angry model and non-angry model. The classification decision is made by comparing the difference between the likelihoods with a threshold.

Language model train set lexical analysis for CCD is given in Table 5.1. The training data is imbalanced with non-angry model containing at least two times that of distinct units in the angry model. This imbalance creates a disadvantage during likelihood calculations for the angry model. In order to compensate for this "add delta smoothing" between angry and non-angry language models is applied. Delta value of 0.25 is found to be optimum in terms of classification accuracies in the train set.

Table 5.2. Percentage of OOV for different language models for CCD.

	Angry Model	Non-angry Model
word	29.85%	7.85%
stem	24.35%	6.39%
stem+ending	23.80%	6.24%

The difference in out of vocabulary percentages on CCD for different language models implemented are presented in Table 5.2. Call center conversations use limited vocabulary and sentence forms. On the contrary, for a general purpose Language model trained from a large Turkish corpus with stem+ending model OOV rate is 2.5% [27].

5.4. Decision Fusion

System combination at the decision level is implemented in order to avoid dimensionality problems. A multi layer perceptron (MLP) is trained for this purpose using neural network toolbox of Matlab. Training set utterance scores for SVM, GMM and LM classifiers are normalized and mapped to [-1 1] range where higher values imply higher probabilities of being angry. These scores are three inputs of MLP and the expected labels are the desired output. Using all of the utterances in the non-angry train set created a bias towards this class. To remove this bias, non-angry utterances as many as angry utterances are included for training. A two layer MLP having three neurons in first layer and a single neuron in second layer applying gradient descent backpropagation training algorithm is found to be optimum for decision merging.

6. EXPERIMENTS

6.1. Experiments with Call Center Data

GMM classifier performances for different number of mixtures are given in Figure 6.1. The performance of the GMM classifier flattens out after 16 mixtures.

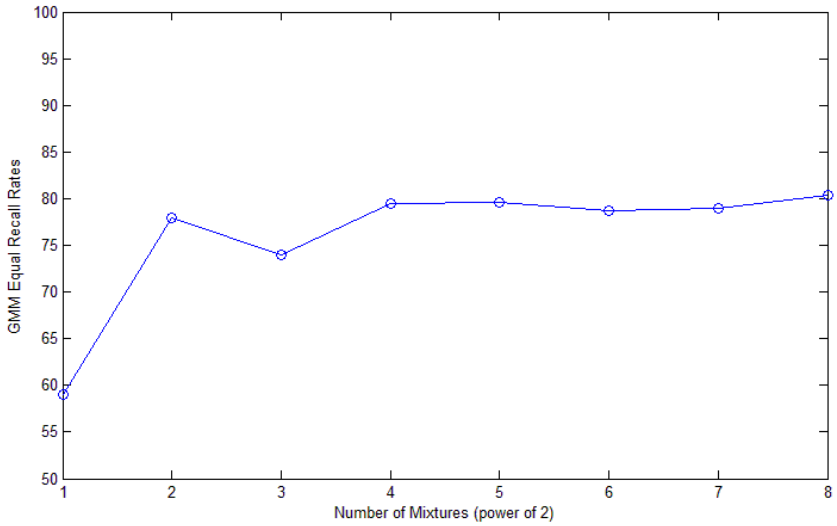

Figure 6.1. Equal recall rates for GMM classifier with different number of mixtures on CCD.

Results for LM classifiers with different modelling units are given in Figure 6.2. When we compare the equal recall rates for anger and non-anger classes, word, stem-only, and stem+ending models resulted in 70.5%, 69.3% and 72.0% respectively. Therefore we can conclude that stem+ending based model performs better than the two other methods.

Language model anger recall values did not increase above 0.74 while non-anger values are above chance level. We can conclude that some of the utterances in the test set labeled as angry, does not contain any linguistic information which can favor angry model over non-angry model. In other words speakers may sometimes speak completely with neutral content even though they are emotionally disturbed. Therefore acoustic information is necessary for correct classification of these utterances. In order

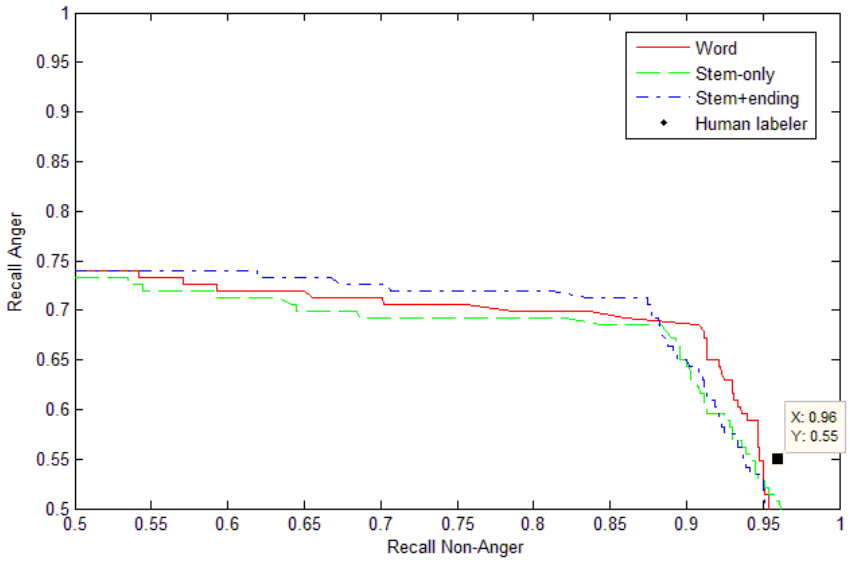

Figure 6.2. Results of different LM classifiers and human labeler performance on CCD.

to confirm this, test set is labeled by a human subject by only reading the transcriptions. An anger recall value of 55% and non-anger recall value of 96% could be achieved.

In Turkish, polite speaking style differs from impolite speaking style mostly by usage of different suffixes. An example for this difference is given in Table 6.1 for three words with two suffixes. The English meanings of words are given in parentheses. Since angry people tend to use less polite speaking style this is reflected in the choice of endings. Therefore removing endings results in degradation of the performance. Also frequencies of semantically similar endings are increased for stem+ending model which explains the performance superiority.

Table 6.1. Turkish stemming example.

Impolite Style	Polite Style
bırak+ın (leave it)	bırak+ınız (please leave it)
çıkart+ın (remove it)	çıkart+ınız (please remove it)
kapat+ın (close it)	kapat+ınız (please close it)

6.2. Comparison of Classifiers for Call Center Data

Results of different classifiers are given in Figure 6.3 as anger recall vs non-anger recall curves. Error rates for various operating points are calculated by applying a threshold to the differences between angry model scores and non-angry model scores.

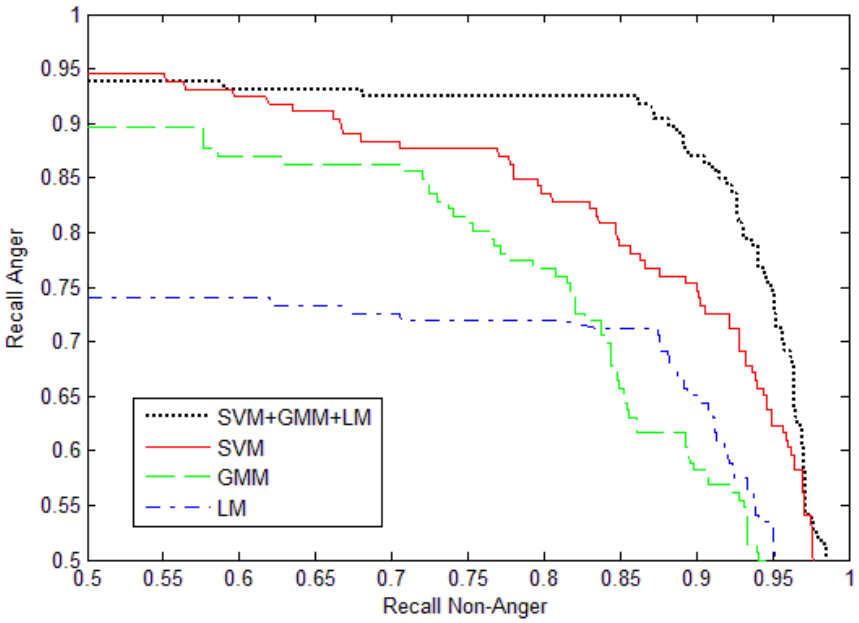

Figure 6.3. Results of different classifiers on CCD.

To measure the similarity between classifiers Q statistic is calculated as in [12].

$$Q = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}} \tag{6.1}$$

where N_{11} is the number of both classifiers being correct, N_{10} is the number of first classifier being correct and second classifier wrong, N_{01} is the number of first classifier being wrong and second classifier correct, N_{00} is the number of both classifiers being wrong. Absolute value of Q being closer to 0 means the classifiers are independent whereas higher values indicate higher similarities between the two classifiers.

The pairwise Q values for SVM, GMM, and LM classifiers are given in Table 6.2. SVM and GMM classifiers have a relatively high correlation as they both model acoustic

Table 6.2. Q statistic for classifiers on CCD.

Q(SVM, GMM)	0.63
Q(SVM, LM)	0.05
Q(GMM, LM)	-0.16

characteristics. The Q values between LM classifier and others are lower which means linguistic channel contains complementary information to acoustic parameters which is expected. Additionally negative Q value with GMM method indicate that correct classification is done for different samples of the test set[37].

6.3. Experiments with FAU Aibo Data

GMM classifier performances for different number of mixtures are given in Figure 6.4. We observe that increasing the number of mixtures used does not increase the performance considerably.

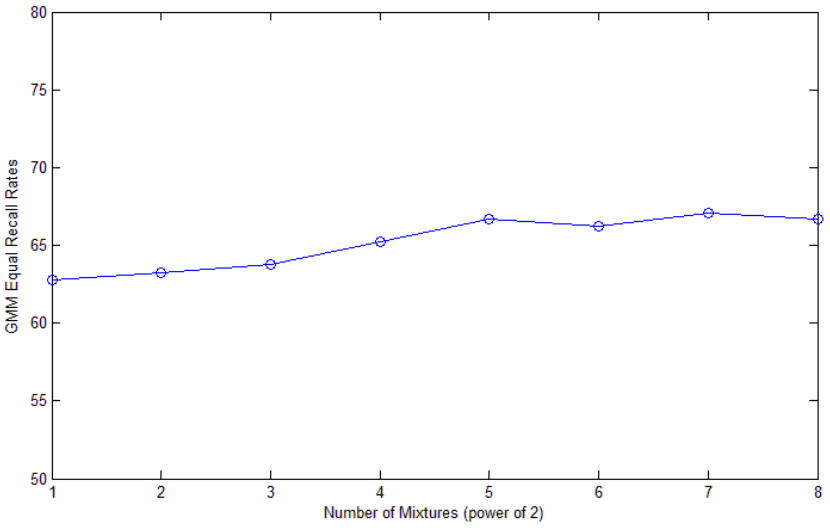

Figure 6.4. Equal recall rates for GMM classifier with different number of mixtures on FAD.

Results for LM classifiers with different modeling units are given in Figure 6.5. When equal recall rates are considered the word model performs the best with 57% accuracy.

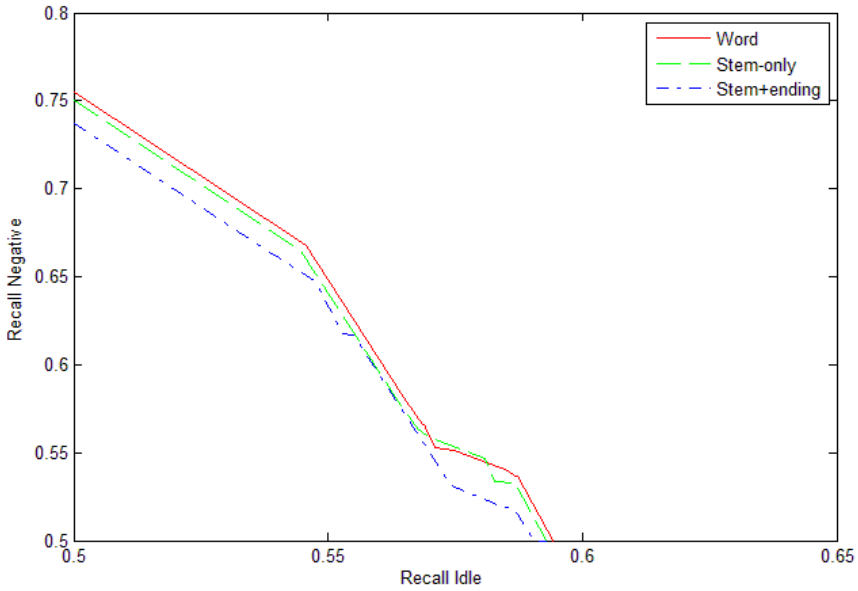

Figure 6.5. Results of different LM classifiers on FAD.

6.4. Comparison of Classifiers for FAU Aibo Data

Results of different classifiers on Fau Aibo dataset are given in Figure 6.6. GMM classifier performed best with 68% accuracy while SVM classifier achieved 66% correct detection. LM classifier performed worse than the other two by making correct decision for 57% of the test set.

The studies on FAD generally report the system performance with AU and WA recall rates. For selecting a threshold, UA rates for different operation points are calculated and plotted in Figure 6.7. A threshold which maximizes the UA rate on train set is chosen. Using this threshold, 68.42% and 63.90% recall rates are achieved on test set for UA and WA respectively.

The pairwise Q values for the three classifiers are given in Table 6.3. Observed Q values are similar to those on call center dataset. There is a high correlation between SVM and GMM classifiers while LM classifier decisions are divergent.

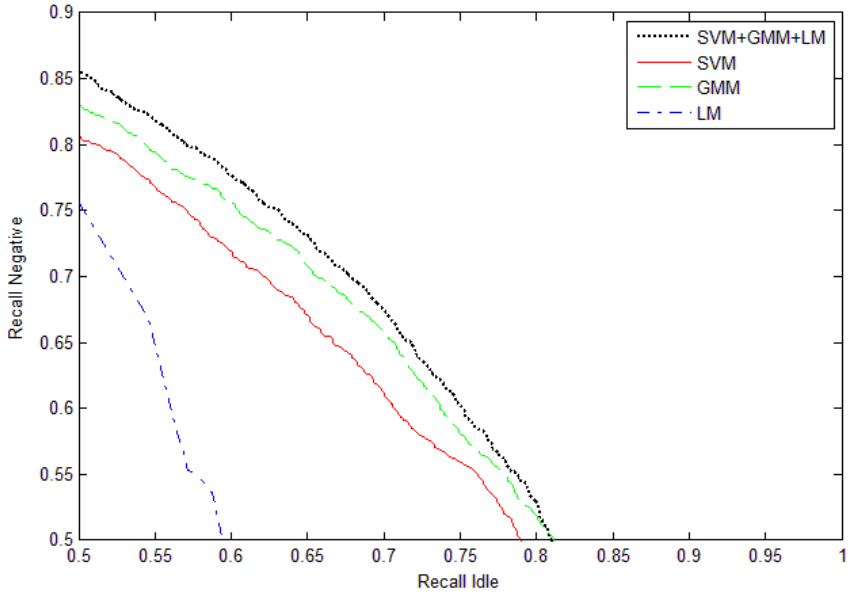

Figure 6.6. Results of different classifiers on FAD.

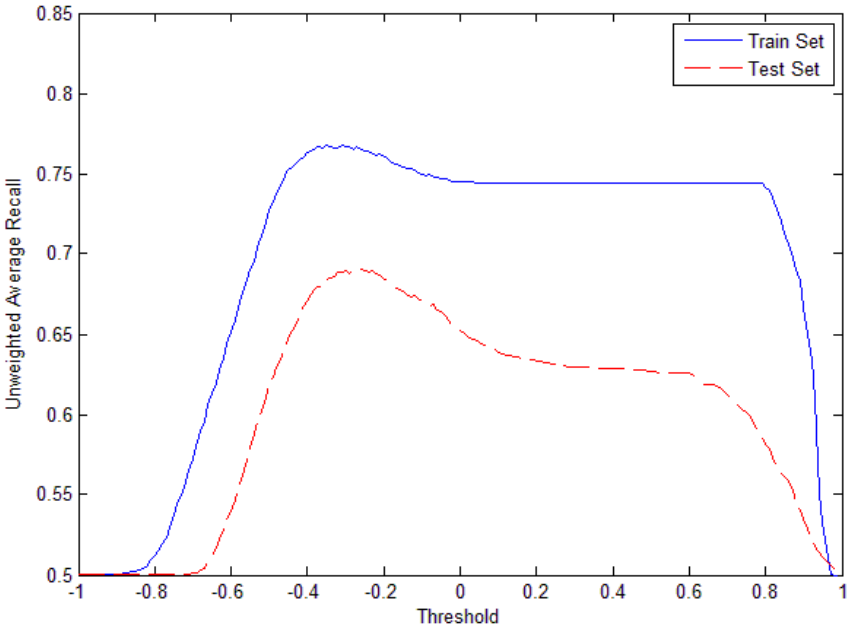

Figure 6.7. UA recall rates vs threshold for train and test sets on FAD.

Table 6.3. Q statistic for classifiers on FAD.

Q(SVM, GMM)	0.77
Q(SVM, LM)	0.17
Q(GMM, LM)	0.41

6.5. Comparison of Results For The Two Datasets

The CCD is formed of agent client recordings whereas FAD consists of utterances directed to a robot. Therefore the content is more restricted to a set of commands for FAD. As a result we have observed that when only linguistic information is used the classification performance on CCD was higher.

Best performing LM is based on stem+ending for CCD and word for FAD. This can be attributed to the differences of Turkish and German language. Turkish is an agglutinative language whereas German is a fusional language. We have used morphs to split words into subword units. Morphs are statistical subparts of words and they are not same as grammatical morphemes. Therefore during the morphological analysis it is more likely to obtain meaningful parts for agglutinative languages when compared with fusional languages.

Decision merging improved the results more for CCD than FAD. The Q value between SVM and GMM classifiers was lower on CCD. In addition to this lower correlation between acoustic classifiers the higher language model performance for CCD resulted in higher increase of accuracy during decision merging phase.

7. CONCLUSION

In this thesis different sources of information are utilized to recognize emotions on two spontaneous datasets. First dataset is composed of Turkish real-life human-human call center data (CCD) while the second one is formed by German speech of children playing with a pet robot, Aibo (FAD). Binary classification was aimed for both databases. On CCD angry vs non-angry emotion classes are considered whereas on FAD negative vs idle emotion classes are taken into account.

Analysis of pitch and energy contours of utterances from different emotion classes revealed that there were substantial statistical differences including range, variations and variations of slopes. When spectograms of same words from negative and idle classes are compared it is observed that higher frequencies were boosted for the former class. Relative frequencies of words are compared in order to detect words correlated to each emotion class. It is found that words “stop”, “no” and “cancel” which indicate disapproval and negative attitude were more frequent in angry and negative classes. Words representing content and affirmation such as “yes”, “okay” and “good” were more frequent in non-angry and idle classes.

For extracting acoustic information we have implemented a SVM classifier with utterance level features and a GMM classifier with frame level features. SVM classifier with utterance level features outperformed GMM with frame level features for CCD. The CCD contains noisy utterances which are recorded through various telephones. Features extracted for SVM classification are more robust to these conditions resulting in a better performance. Conversely on FAD GMM achieved better accuracy than SVM because of the noiseless high quality recordings.

In order to extract linguistic information from utterances Unigram LM classifiers are implemented based on words, stem-only and stem+ending units. On CCD Stem+ending based model achieved accuracies higher than word based model because of the level of politeness conveyed through suffixes in Turkish. Additionally we ob-

served that LM classifier could categorize utterances correctly which are misclassified by acoustic classifiers. As a result, after merging the scores of three classifiers by an MLP recognition accuracies of anger and non-anger increased considerably.

Anger recall values for LM classifier on CCD did not increase beyond 74% while the non-anger recall values were above chance level. This indicates that some of the utterances in the test set are not classifiable using only linguistic channel. Therefore acoustic information is necessary for correct classification of these utterances.

In [22] it is observed that stemming enhances lexical model performance on French human-human call center data. This is contrary to our findings on both datasets investigated. For CCD this can be attributed to the different morphological structures of Turkish and French. Also level of politeness conveyed through suffixes in Turkish can be another reason. German is a fusional language like French. However, the datasets are from different domains. Additionally in our experiments we have used Morphessor to split words into subword parts in a statistical manner. So the resulting stems sometimes differ from grammatical stems for German which may explain the decrease in performance on FAD.

Human human dialog data is rarely investigated in literature in terms of emotion recognition. A similar study to this one is presented in [3], which is on human-human call center data classifying between negative and positive emotions. We have achieved comparable accuracies with SVM classifier using acoustic features.

The studies on FAD are summarized in [38]. Similar to our implementation acoustic and linguistic classifiers are built and mostly MFCC related features are employed. For the two class problem baseline UA was 67.7%. The reported UA values are ranging from 66.4% to 70.3%. We have achieved an UA of 68.42% which is above the baseline score and below the best result reported.

Future work includes considering the whole dialog while making a decision on utterance level as well as dialog level. Also the effect of using automatic speech recog-

nizer (ASR) hypothesis instead of manual transcriptions for language modeling will be investigated. Additionally considering grammatical morphemes instead of statistical morphs can enhance the performances of LM classification.

REFERENCES

1. Pan, Y. C., M. X. Xu, L. Q. Liu, P. F. Jia "Emotion-detecting Based Model Selection for Emotional Speech Recognition", in Proceedings of the Multiconference on Computational Engineering in Systems Applications, Beijing, 2006.

2. Steidl, S., M. Levit, A. Batliner, E. Nöth and H. Niemann, "Off All Things the Measure is Man Automatic Classification of Emotions and Inter-labeler Consistency", Proceeding of the International Conference on Acoustics, Speech, and Signal Processing, 2005.

3. Vidrascu, L. and L. Devillers, "Detection of Real-life Emotions in Call Centers," in Proceedings of Interspeech, 2005.

4. Luengo, I., E. Navas, I. Hernáez, and J. Sánchez, "Automatic Emotion Recognition Using Prosodic Parameters", In Proceedings of Interspeech, pp. 493-496, 2005.

5. Morales-Perez, M., J. Echeverry-Correa, A. Orozco- Gutierrez and G. Castellanos-Dominguez, "Feature Extraction of Speech Signals in Emotion Identification", IEEE International Conference of the Engineering in Medicine and Biology Society. Vancouver, Canada, 2008.

6. Fu, L., X. Mao, L. Chen, "Speaker Independent Emotion Recognition Using HMMs Fusion System with Relative Features", in Intelligent Networks and Intelligent Systems, 2008.

7. Yacoub, S., S. Simske, X. Lin and J. Burns, "Recognition of Emotions in Interactive Voice Response Systems", In Eurospeech, 729-732, 2003.

8. Dellaert, F., T. Polzin and A. Waibel, "Recognizing Emotion in Speech." In Proceedings of International Conference on Spoken Language Processing, Philadelphia, 3:1970-1973, 1996.

9. Lee, C., M., S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan, "Emotion Recognition Based on Phoneme Classes", Proceedings of the International Conference on Spoken Language Processing, 2004.

10. Yaslan, S. and B. Günsel, "Emotion Recognition from Digital Audio Signals", Signal Processing and Communications Applications Conference, 2005.

11. Vogt, T., and E. Andre,"Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition", Proceedings of International Conference on Multimedia and Expo, Amsterdam, Holland, 2005.

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

13. Burkhardt, F., T. Polzehl, J. Stegmann, F. Metze, and R. Huber, "Detecting Real Life Anger", In Proceedings International Conference on Acoustics, Speech, and Signal Processing, Taipei, Taiwan, 2009.

14. Inanoglu, Z., R. Caneel, "Emotive Alert: HMM-Based Emotion Detection in Voicemail Messages", In Proceedings of Intelligent User Interfaces, 251-253, 2005.

15. Vidrascu, L., and L. Devillers, "Five Emotion Classes Detection in Real-world Call Center Data: the Use of Various Types of Paralinguistic Features", Workshop Paraling07, 2007.

16. Neiberg, D., K. Elenius and K. Laskowski, "Emotion Recognition in Spontaneous Speech Using GMMs," In Proceedings of International Conference on Spoken Language Processing, Pittsburgh, pp. 809812, 2006.

17. Schuller, B, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in Interspeech, Isca, Brighton, UK, 2009.

18. Polzehl, T., S. Sundaram, H. Ketabdar, M. Wagner, and F. Metze, "Emotion Classification in Childrens Speech Using Fusion Acoustic and Linguistic Features",

Proceedings of Interspeech, Brighton, UK, 2009.

19. Bozkurt, E., E. Erzin, C. Eroglu Erdem, A, T. Erdem "Improving Automatic Emotion Recognition from Speech Signals", Proceedings of Interspeech, Brighton, UK, 2009.

20. Schuller, B., F. Metze, S. Steidl, A. Batliner, F. Eyben, T. Polzehl, "Late Fusion of Individual Engines for Improved Recognition of Negative Emotion in Speech - Learning vs. Democratic Vote." In Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pp. 5230-5233, 2010.

21. Schuller, B., D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards More Reality in the Recognition of Emotional Speech.", In IEEE, editor, In Proceedings of International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 941-944, Honolulu, Hawaii, USA, 2007.

22. Devillers, L., and L. Vidrascu, "Real-life Emotions Detection with Lexical and Paralinguistic Cues on Human-human Call Center Dialogs", Interspeech 2006.

23. Metze, F., A. Batliner, F. Eyben, T. Polzehl, B. Schuller, and S. Steidl, "Emotion Recognition Using Imperfect Speech Recognition", Proceedings of Interspeech, pages 478-481, Makuhari, Japan, 2010.

24. Müller, R., B. Schuller, G. Rigoll "Enhanced Robustness in Speech Emotion Recognition Combining Acoustic and Semantic Analyses", In Proceedings of Workshop From Signal To Signs of Emotion and Vice Versa, EU-IST Network of Excellence Humaine, Santorini, Greece, September 2004.

25. Schuller, B., G. Rigoll, and M. Lang, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture," in Proceedings of International Conference on Acoustics, Speech, and Signal Processing, pp. 577580, 2004.

26. Shafran, I. and M. Mohri, "A Comparison of Classifiers for Detecting Emotion from Speech," Proc. of International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, Mar 19-23, 2005.

27. Arisoy, E., M. Saraclar, T. Hirsimäki, J. Pylkkönen, T. Alumäe, H. Sak, Fr. Mihelic, J. Zibert, (eds.), "Statistical Language Modeling for Automatic Speech Recognition of Agglutinative Languages", Speech Recognition : Technologies and Applications, Ch. 10, pp. 194–204, 2008.

28. Ekmekçioğlu, F. Ç. and P. Willet. "Effectiveness of Stemming for Turkish Text Retrieval", Program, 34(2):195200, April 2000.

29. Rong, J., Y.-P. P. Chen, M. Chowdhury, and L. Gang;. "Acoustic Features Extraction for Emotion Recognition." In Proceedings of 6th International Conference on Computer and Information Science, Melbourne, Australia, pages 419424, July 2007.

30. Steidl, S., "Automatic Classification of Emotion-related User States in Spontaneous Children's Speech", Ph.D. Thesis, Logos, Berlin, Germany, 2009.

31. Batliner, A., S. Steidl, C. Hacker, and E. Nöth, "Private Emotions vs. Social Interaction a Data-driven Approach Towards Analysing Emotion in Speech," in User Modeling and User-adapted Interaction, vol. 18, pp. 175206, 2008.

32. Chang, Chih-Chung and Lin, Chih-Jen., "LIBSVM: a Library for Support Vector Machines", 2001.

33. Talkin, D., "A Robust Algorithm for Pitch Tracking (RAPT)," in Speech Coding and Synthesis (Elsevier, ed.), pp. 495518, 1995.

34. Young, S., D. Ollason, V. Valtchev and P. Woodland, "The HTK Book" (for HTK Version 3.2), Entropic Cambridge Research Laboratory, 2002.

35. Blouet, R., C. Mokbel, H. Mokbel, E. Sánchez Soto, G. Chollet, and H. Greige, "Becars: a Free Software for Speaker Verification", Odyssey, Spain, 2004.

36. Creutz, M. and K. Lagus, "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor", Publications in Computer and Information Science, Report A81, Helsinki University of Technology, March, 2005.

37. Kuncheva, L., and C. Whitaker, "Measure of Diversity in Classifier Ensembles," Machine Learning, vol. 51, pp. 181207, 2003.

38. Schuller, B., A. Batliner, S. Steidl, D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge", Speech Communication, Special Issue Sensing Emotion and Affect - Facing Realism in Speech Processing, to Appear.