Emotions Analysis of Speech for Call Classification

Esraa Ali Hassan[1], Neamat El Gayar [2], Moustafa M. Ghanem[3]
Center for Informatics Science
Nile University
Giza, Egypt
e-mail: esraa.ali@nileu.edu.eg [1], {nelgayar[2] , mghanem[3]}@nileuniversity.edu.eg

Abstract— **Most existing research in the area of emotions recognition has focused on short segments or utterances of speech. In this paper we propose a machine learning system for classifying the overall sentiment of long conversations as being Positive or Negative. Our system has three main phases, first it divides a call into short segments, second it applies machine learning to recognize the emotion for each segment, and finally it learns a binary classifier that takes the recognized emotions of individual segments as features. We investigate different approaches for this final phase by varying how emotions for individual segments are aggregated and also by varying classification model used for the final phase. We present our experimental results and analysis based on a simulated data set collected specifically for this research.**

Keywords:- speech analysis; emotions recognition; classification of calls; machine learning.

I. INTRODUCTION

Automated sentiment analysis, or emotions recognition, from audio recordings can provide a vital and powerful tool to help many organizations make better decisions. For example, companies could use such technology to analyze customer-service and helpdesk conversations or even voice mail [1, 2]. Law enforcement and intelligence organizations could apply the technology to analyze intercepted phone conversations. Public relations firms could employ it to analyze news broadcasts to find coverage of clients [3].

Automatic emotions recognition from speech refers to the use of various methods to analyze vocal behavior as a marker of affect (e.g., emotions, moods, and stress), focusing on the nonverbal aspects of speech [4]. The basic assumption is that there is a set of objectively measurable voice parameters that reflect the affective state a person is currently experiencing (or expressing for strategic purposes in social interaction). In this context, automatic emotions recognition from audio recordings poses several interesting and new theoretical challenges including; identifying relevant features of the audio signal and creating robust statistical and learning methods for automatically discovering the patterns in these features.

In this research we focus on the task of binary classification of call recordings into Positive and Negative sentiment. Our motivation is developing a system that can be used in call centers to detect whether customers are happy with the level of service offered by a call center. We propose a model suited for long conversations. Our approach is based on analyzing the acoustic features of the dialogue in the recording, recognizing different emotions for short segments in the call, and relating the recognized emotions to the final conversation label. Our hypothesis is that by extracting the emotions throughout the call, the machine learning techniques will be able to identify, and automatically learn patterns of emotions to classify the call as being positive or negative.

The remainder of this paper is organized as follows. Section II describes related background to automated speech sentiment analysis and emotions recognition. Section III describes the proposed system. Section IV presents our experiments and results. Finally in Section V the paper is summarized and concluded.

II. BACKGROUND

A typical speech emotion recognition system consists of three principal stages: signal pre-processing and segmentation, followed by feature extraction and calculation, then finally the classification stage [5]. The first stage involves acoustic preprocessing like signal filtering, as well as segmenting the input signal into meaningful units. Feature extraction and calculation is concerned with identifying relevant features of the acoustic signal with respect to emotions. Finally, classification maps feature vectors onto emotion classes through learning by examples.

The goal of the audio segmentation is to segment a speech signal into units that are representative for emotions [16]. These are usually linguistically motivated middle-length time intervals such as words or utterances. Generally speaking, a good emotion unit has to fulfill certain requirements. In particular, it should be (1) long enough to reliably calculate features by means of statistical functions

978-1-4244-8136-1/10/$26.00 ©2010 IEEE

and (2) short enough to guarantee stable acoustic properties with respect to emotions within a segment [5].

There has been an increasing amount of work in identifying the acoustic features that vary with the speaker's emotional state. All studies in the field point to the pitch (the fundamental frequency) as the main vocal cue for emotions recognition [6]. Additional acoustic variables contributing to vocal emotion signaling are: vocal energy, frequency spectral features, and formants (usually only one or two first formants (F1, F2) are considered), and temporal features (speech rate and pausing). Another approach to feature extraction is to enrich the set of features by considering some derivative features such as LPC (linear predictive coding) parameters of signal or features of the smoothed pitch contour and its derivatives [20].

As the research in emotion recognition started since the 80s [20] almost all well known classification techniques were used in the recognition process like: K-nearest neighbors [8], Neural Networks [8, 10], Decision Trees [11], Support Vector Machine [12, 13], Naive Bayes [5], Multiple Classifier Systems [14, 15], Echo State networks [7] which is a type of recurrent neural networks, and ensemble of neural networks [9].

The majority of the work done in the area of emotion recognition is concerned with finding the finest features to extract, and the most accurate classifiers to be used in recognizing the emotion for short segments or utterances of speech. In this work we contribute to the field by introducing a new system that uses the different emotions recognized from the entire conversation in classifying the call.

III. PROPOSED SYSTEM

Our proposed model consists of three main phases as shown in figure 1, signal segmentation, emotions recognition, and calls classification. The system starts by segmenting the call into small homogeneous segments, then it recognizes the emotion of each segment in the second phase, finally in the third phase the system analyzes the emotions recognized throughout the call to classify it as 'Positive' or 'Negative'. In the following three subsections we describe each component in detail.

A. Segmentation

The goal of the segmentation process is to produce a sequence of discrete utterances with particular characteristics remaining constant within the segment. The segmentation model is adopted from [16], in which the segmentation algorithm locates times in the audio stream where there is a change in the acoustic class. The process starts by extracting *Mel Frequency cepstral coefficients* (MFCC) features, then detecting the areas where there is a change in the energy using a sliding window over the audio stream and a Kullback Leibler (KL) distance metric. Finally, after marking the initial start and end of each segment, it searches around them for the low energy parts to segment this portion.

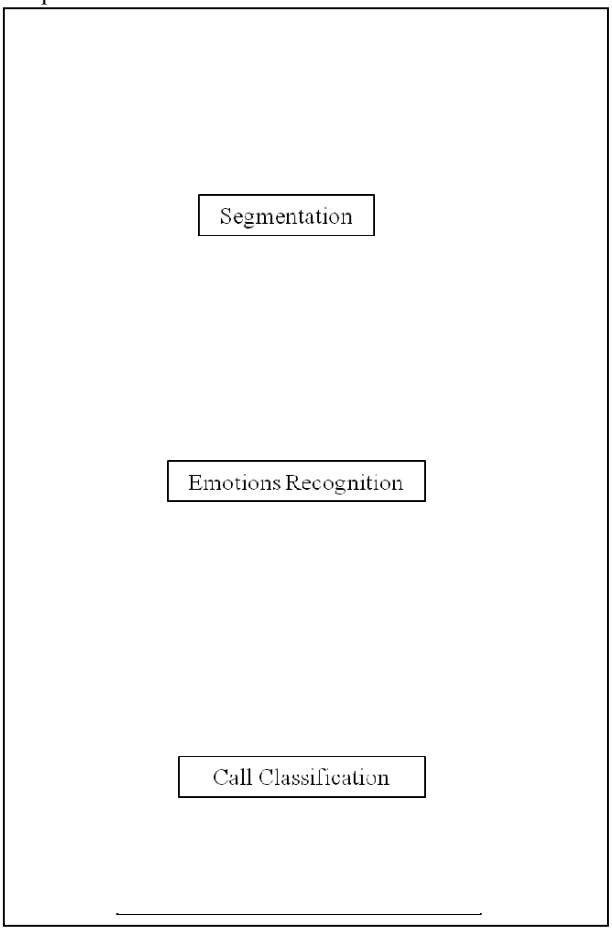

Figure 1: System Block Diagram

B. Emotions recognition

After segmenting the call to shorter segments with stable acoustic properties that ensure a constant emotional state within this segment, the emotions recognition module labels each segment with one of seven emotions namely (*happy, neutral, boredom, disgust, fear, sadness, anger*) [21]. The module uses the acoustic features extracted to classify the segment to one of seven emotions.

The features extracted in this phase are the pitch, formants (first and second formants only), energy, 12 *Mel frequency*

cepstral coefficients (MFCC) [24], and *linear predictive coding coefficients* (LPCC) [15]. We compute the maximum, minimum, average, standard deviation and variance for each extracted feature along with the original signal. Thus we have a total of 90 features that serve as input to the classification module.

For the classification process a Multilayer perceptron technique is used. A multilayer perceptron is an artificial neural network structure estimator that can be used for classification and regression and it can distinguish data that is not linearly separable [22].

C. Call Classification

The ultimate target of the call classification phase is to employ the sequence of emotions recognized throughout the dialogue, calculate features from this sequence, and finally use these features to classify the call as being positive or negative. As follows we describe four approaches applied for extracting features from recognized emotions of the call to classify the conversation. Experimental results are presented in the next section.

a) *Normalized Emotions Scores (NES) :*

The first approach takes all the emotions of the call into account; this is achieved by counting the occurrences of each emotion throughout the call, and then normalizing the count by dividing it by the total number of segments of the call.

This approach results in having a vector of 7 features - one for each emotion label. This vector is then used as an input to the classifier. Figure 2 summarizes this approach.

b) *Normalized Emotion Category Scores (NECS):*

In this approach we aggregate the emotions into 2 categories: {*Positive, and Negative emotions*}, count the occurrences of each category in the conversation, normalize the count and use the final vector as an input to the classifier.

As we can see from Figure 3; The *Positive* category includes the *happy*, *neutral*, *boredom*, and *fear* emotions, while the *Negative* category contains the *sadness*, *anger*, and *disgust* emotions.

c) *Last-K Emotions (LKE) :*

Considering the importance of the last segments in deciding on the call final label, the fourth approach takes only the later segments of the call into account and ignores all the earlier ones.

The idea behind this approach is the hypothesis that usually the last part of the call indicates to a bigger extent whether this call has terminated in a 'satisfactory' or 'unsatisfactory' manner. Hence this method labels a conversation as being "positive" or "negative", by focusing on detecting the emotions in the last segments of the call. For this approach the user needs to specify the ***K*** last segments of the call to be taken into account while labeling the conversation.

Inputs
- **N =** number of segments in the call
- ***S*****=** [S_0,S_1, ... ,S_N], vector contains segments of the call

Algorithm
- **initialize** *HappyCount, SadnessCount, FearCount, BoredomCount, NeutralCount, AngryCount, and DisgustCount* **to 0**
- **For each** *segment* **in** *S*
- **Switch (** Emotion(*segment*) **)**
- **Case** (*"Happy"*) : **increment** *HappyCount*
- **Case** (*"Sadness"*): **increment** *SadnessCount*
- **Case** (*"Fear"*) : **increment** *FearCount*
- **Case** (*"Boredom"*): **increment** *BoredomCount*
- **Case** (*"Neutral"*): **increment** *NeutralCount*
- **Case** (*"Angry"*): **increment** *AngryCount*
- **Case** (*"Disgust"*): **increment** *DisgustCount*
- **End Switch**
- **End loop**
- **Normalize** *HappyCount , SadnessCount, FearCount, BoredomCount, NeutralCount, AngryCount, and DisgustCount* **divide by** N
- **Return feature_vector = [** *HappyCount , SadnessCount, FearCount, BoredomCount, NeutralCount, AngryCount, and DisgustCount* **]**

Figure 2: Normalized Emotions Scores Algorithm

Inputs
- **N =** number of segments in the call
- ***S*****=** [S_0,S_1, ... ,S_N], vector contains segments of the call

Algorithm
- **initialize** *PositiveCount, NegativeCount and NeutralCount* **to 0**
- **For each** *segment* **in** *S*
- **Switch (** Emotion(*segment*) **)**
- **Case** (*"Happy"* ***or*** *"Fear"*
  or *"Neutral"* ***or*** *"Boredom"*):
- **increment** *PositiveCount*
- **Case** (*"Sadness", "Disgust",* ***or*** *"Angry"*) :
- **increment** *NegativeCount*
- **End Switch**
- **End loop**
- **Normalize** *PositiveCount, and NegativeCount* **divide by** N
- **Return feature_vector = [***PositiveCount, and NegativeCount* **]**

Figure 3: Normalized Emotion Category Scores Algorithm

d) *Dominant Emotions across M Intervals (DEMI):*

A more sophisticated approach to follow is to summarize the emotions in the conversation into ***M*** equal portions. Each part is then labeled after the most frequent emotion in its segments.

Inputs

- **No_Seg =** number of segments in the call
- **M**= no of intervals to summarize the call in.
- ***S*****=** [S_0,S_1, ... ,S_N], vector contains segments of the call ordered by the start time of each segment.

Algorithm

- **Calculate** *Step* = **No_Seg** / **M**
- **initialize** *i* ***and*** *j* ***to 0***
- **Create** *SummarizedVector* **to be** a vector of length =M , to hold the summary of emotions.
- **Repeat until** *j* = **M**
- **Set** *SummarizedVector[j]* = the most frequent emotion in the sub vector of ***S,*** which starts **from S[** *i***] to S[** *i*+ *Step***]**
- **Update** *i* **to be** *i* + *Step*
- **Update** *j* **to be** *j* +*1*
- **End loop**
-
- **Return feature_vector =** *SummarizedVector*

Figure 4: Dominant Emotions across M Intervals Approach Algorithm

The rationale of this approach is to consider all the emotions in the call, and in the same time, maintain the pattern of how the emotions change from the start to the end of the conversation. As the number of the intervals ***M*** increases, more detailed information can be captured about the emotions and the way they vary throughout the call.

IV. EXPIREMENTS AND RESULTS

In this section we present the details of our experiments. In particular we start by describing the data collection process, followed by classifier training and evaluation, along with results' discussion.

A. Data Collection

To train the emotions recognition classifier, the Berlin database of emotional speech was used [21]. It contains about 500 utterances spoken by actors in a *happy, angry, sadness, fearful, bored, disgusted* and *neutral* emotional state. Ten different actors and ten different texts are used in recoding the dataset.

To evaluate the overall system performance another dataset was collected especially for this purpose. The dataset used in our experiments consists of 39 calls simulating real agent/customer conversations in a call center given predefined scripts. Ten persons (males and females) participated with different combinations in recording 12 scenarios in a controlled environment.

The twelve predefined scripts are divided into 22 positive and 17 negative scenarios. The calls which end with both conversation parties satisfied are labeled *"Positive"*, the scenarios that end with both of the participants not satisfied are labeled "*Negative*".

B. Expiremental Results

1) Preprocessing and Segmentation

As a preprocessing step; all the audio files are converted to PCM signed 16 bit, mono format, with 16 KHz sampling rate. The DC (distortion) bias is removed from the original signal by subtracting the mean value from the signal. For the segmentation the Sphinx4 [10] tool is used for the computation of features from the signal, the features are composed of 13 MFCCs with coefficient C_0 as energy. The open source tool LIUM_Spkdiarization [18] is used for the segmentation. The distance metric used is Kullback Leibler (KL), with a threshold value of 3, and window size of .5 second.

2) Emotion Recognition

In the emotions recognition phase, the features are extracted for both the testing and training datasets using COLEA; a MATLAB toolbox for speech analysis [22]. The WEKA (Waikato Environment for Knowledge Analysis) [19] data mining toolbox is used to train the multilayer perceptron classifier. The Multilayer perceptron learning rate is 0.3, number of hidden layers is 5, and momentum value is 0.2. As mentioned earlier, the classifier is trained offline using Berlin database for emotional speech.

The trained classifier is tested on 50 segments manually labeled by multiple users.The classifier accuracy for recognizing the segments' emotions was about 76%. For our conversations most recognized emotions were found to belong to the category of neutral, happiness, disgust and sadness. Other emotions like fear and boredom were rarely detected.

3) Call classification

In our experiments, we use the generated vector from *NES*, *NECS*,*LKE* and *DEMI* approaches as an input to different classifiers in WEKA for the purpose of comparing their performance.

We compare the performance of the following classifiers: a) the J48 decision tree classifier, which is a java implementation for C4.5 decision tree algorithm, b) the CART decision tree classifier c) the NaiveBayes classifier, d) the KStar classifier which is an instance-based classifier, and finally e) the Multilayer Perceptron classifier [19].

The classifiers are trained and evaluated with 10 folds cross validation. Furthermore the Accuracy and F-measure scores are computed for each classifier. Table 1 summarizes the obtained accuracies, and table 2 shows the F-Measure values.

TABLE I. COMPARING ACCURACY OF DIFFERENT APPROACHES

Approach	J48	CART	Naïve Bayex	K*	MLP
NES	59.33	62.33	47	**67.58**	55.25
NECS	60.58	**64.58**	51.58	65.42	**64.50**
LKE *last 7 segments*	59.83	54.75	48	43.33	35.33
LKE *last 10 segments*	62.58	47.58	56.75	52.17	50.17
LKE *last 15 segments*	58.5	52	**63.33**	53.67	60
DEMI *No of parts = 10*	65.75	56.42	52.08	53.17	51.33
DEMI *No of parts = 15*	**79**	59.83	48.42	60.58	50.75

As we can see from the results, there is no winning classifier for all the approaches. However, the J48 classifier achieves relatively better accuracies and F-measure values with highest f-measures against all the other classifiers. The best result was for the Dominant Emotions across M Intervals approach with J48 classifier.

TABLE II. COMPARING F-MEASURE OF DIFFERENT APPROACHES

Approach	J48	CART	Naïve Bayex	K*	MLP
NES	0.4	0.4	0.21	0.51	0.40
NECS	0.15	0.44	0.29	0.48	0.40
LKE *last 7 segments*	0.61	0.62	0.46	0.38	0.32
LKE *last 10 segments*	0.59	0.5	0.52	0.49	0.46
LKE *last 15 segments*	0.56	0.53	0.61	0.48	0.54
DEMI *No of parts = 10*	0.5	0.18	0.36	0.29	0.32
DEMI *No of parts = 15*	0.67	0.18	0.24	0.33	0.35

On the other hand, the multilayer perceptron has the lowest accuracies among all the used classifiers. For MLP accuracies range between 55.52% and 64.50% which is very poor.

In general the DEMI approach –which summarizes the emotions in the conversation into ***M*** equal portions - could achieve the best results across the different classifiers. This approach reaches 79% on J48 classifier. Experiments show that accuracy improves with larger values for the number of parts. Evidently as the 15 parts gives more detailed information about the call than the 10 parts, this leads to better capturing the emotional patterns throughout the call.

Although the NES and NECS approaches are based on counting the emotions without considering when they appear in the call, they are doing quite well for CART, KStar, and MLP classifiers. It is also interesting to notice that the resultant trees from both of them are quite meaningful; the tree of the first approach is shown in figure 5.

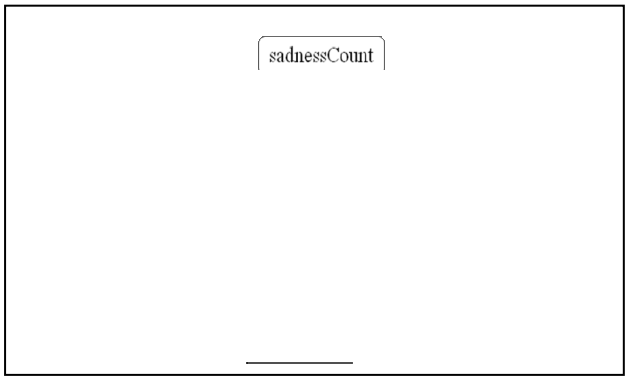

Figure 5: NES Generated Tree

As we can see from the previous figure, when the count of sadness emotions decreases and count of happy emotions increases the call is classified as "Positive". On the other hand when the sadness count is lower than a certain threshold, the fear count decreases, and count of disgust emotions increases the call is classified as "Negative" call, and so on for the rest of the tree.

The last K emotions approach unexpectedly achieves poor accuracy for all the classifiers regardless of the value of K. As mentioned before this study presents preliminary results based on acted data. More conclusive results will be obtained based on data gathered from real conversations.

V. CONCLUSION & FUTURE WORK

In this paper, we propose a system to analyze the emotions in calls based on non-verbal aspects of speech. Preliminary experiments are conducted on an acted dataset and show promising results.

We present four different approaches to analyze the emotions throughout the call. Experiments reveal that the *Dominant Emotions across M Intervals* approach obtains the best results with 79% accuracy. We also show that even the simplest approaches in extracting emotions from conversation are giving well interpretable classification rules.

However, for more conclusive results several issues need to be addressed; yet the error accumulation between system's phases needs more investigations. On the other hand the system's results seem to be highly dependent on the dataset used. More experiments should be carried out on real datasets to verify the usefulness of the different approaches proposed.

In addition to that, enhancing the different phases' accuracies will result in an overall improvement of the entire system. For the segmentation, a speaker clustering step could be added after the segmentation module. The

speaker clustering identifies the speakers' turns, and this extra information can be used as a new feature for the emotion analysis phase.

The emotion recognition accuracy can be enhanced by extracting more features. Since there is no common agreement on the best features and also the choice of features is to a great extent data dependent; it is important to optimize the choice of the feature set using feature selection techniques.

For the emotion analysis approaches, we are in the process of collecting real dataset for further experiments. In addition, we intend to exploit time series classification techniques to give better indication about the change of emotions during the conversation.

Moreover, we are currently developing a system for speech analysis that integrates the emotions recognition module with analysis of text using sentiment analysis. By combining these two approaches the system should be able to identify, and automatically learn groups of words and emotions that are associated with positive or negative conversations.

REFERENCES

[1] L. V. Subramaniam, T. Faruquie, S. Ikbal, S. Godbole, and M. Mohania, "Business intelligence from voice of customer," Proc. of ICDE, 2009.

[2] H. Takeuchi, L. Subramaniam, T. Nasukawa, and S. Roy, "Automatic Identification of Important Segments and Expressions for Mining of Business-Oriented Conversations at Contact Centers," Proc. of (EMNLP-CoNLL), 2007.

[3] N. Leavitt, "Let's hear it for audio mining," IEEE Comput. Mag., vol. 35, no. 10, pp. 23–25, Oct. 2002.

[4] Juslin, P., and Scherer, K., " Vocal expression of affect," In J. Harrigan and R. Rosenthal and K. Scherer (Ed.), The New Handbook of Methods in Nonverbal Behavior Research (pp. 65-135), Oxford: Oxford University Press, 2005.

[5] Vogt, E André, and J Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation," In Peter, C. and Beale, R., editors, Affect and Emotion in Human-Computer Interaction, volume 4868 of LNCS, 2007.

[6] L. Vidrascu, and L. Devillers, "Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features," In ParaLing'07, 16th International Conference on Phonetic Sciences, Saarbrcken, 2007.

[7] Scherer, M. Oubbati, F. Schwenker, and G. Palm, "Real-time emotion recognition from speech using echo state networks," Proc. Artificial Neural Networks and Pattern Recognition workshop, 2008.

[8] V Petrushin, "Emotion In Speech: Recognition And Application To Call Centers," Artificial Neural Network In Engineering, 1999.

[9] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilan, A. Batliner, N. Amir, and K. Karpousis, "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," in Affective Computing and Intelligent Interaction, A. Paiva, R. Prada, and R. W. Picard, Eds. Berlin-Heidelberg: Springer, 2007, pp. 488–500.

[10] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source frame- work for speech recognition," Tech. Rep., Sun Microsystems Inc., 2004.

[11] J Cichosz, K Ślot, "Emotion recognition in speech signal using emotion extracting binary decision trees," Doctoral Consortium. ACII, 2007.

[12] L Devillers, and L Vidrascu, "Real-Life Emotion Recognition in Speech," In: M¨uller, C. (ed.) Speaker Classification II. LNCS(LNAI), vol. 4441, Springer,2007.

[13] OW. Kwon, K. Chan, J. Hao, and TW. Lee, "Emotion recognition by speech signals," In Proc. of Eurospeech 2003, 2003.

[14] S. Scherer, F. Schwenker, and G. Palm, "Classifier fusion for emotion recognition from speech," in 3rd IET International Conference on Intelligent Environments (IE 07), 2007.

[15] S Scherer, F Schwenker, and G Palm, 'Emotion recognition from speech using multi-classifier systems and RBF-ensembles', in Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, 49–70, Springer,2008.

[16] H. Meinedo, J. Neto,, "Audio segmentation, classification and clustering in a broadcast news task," Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, ,2003.

[17] Q. Jin,K. Laskowski,T. Schultz, and A. Waibel, "Speaker segmentation and clustering in meetings," NIST Spring Rich Transcription Evaluation Workshop,2004.

[18] S. Meignier, and T. Merlin, "Lium SpkDiarization: An open source toolkit for diarization," In CMU Sphinx Users and Developers Workshop, 2010.

[19] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann, San Francisco, 2 edition, 2005.

[20] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication, vol. 48(9), pp. 1162–1181, 2006.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in Proc. Interspeech,2005.

[22] E. Alpaydin, Introduction to Machine Learning, MIT Press,2004.

[23] C.M. Lee and S.S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," IEEE Trans. Speech and Audio Processing, vol. 13, no. 2, pp. 293-303, 2005.

[24] S. Molau, M. Pitz, R. Schl¨uter, H. Ney, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum," Proc. Int. Conf. on Acoustic, Speech and Signal Processing, 2001