Combining Semantic and Acoustic Features for Valence and Arousal Recognition in Speech

Seliz Gülsen Karadoğan
Department of Informatics and Mathematical Modelling
Technical University of Denmark
Copenhagen, 2800, Denmark
Email: seka@imm.dtu.dk

Jan Larsen
Department of Informatics and Mathematical Modelling
Technical University of Denmark
Copenhagen, 2800, Denmark
Email: jl@imm.dtu.dk

***Abstract*—The recognition of affect in speech has attracted a lot of interest recently; especially in the area of cognitive and computer sciences. Most of the previous studies focused on the recognition of basic emotions (such as happiness, sadness and anger) using categorical approach. Recently, the focus has been shifting towards dimensional affect recognition based on the idea that emotional states are not independent from one another but related in a systematic manner. In this paper, we design a continuous dimensional speech affect recognition model that combines acoustic and semantic features. We design our own corpus that consists of 59 short movie clips with audio and text in subtitle format, rated by human subjects in arousal and valence (A-V) dimensions. For the acoustic part, we combine many features and use correlation based feature selection and apply support vector regression. For the semantic part, we use the affective norms for English words (ANEW), that are rated also in A-V dimensions, as keywords and apply latent semantics analysis (LSA) on those words and words in the clips to estimate A-V values in the clips. Finally, the results of acoustic and semantic parts are combined. We show that combining semantic and acoustic information for dimensional speech recognition improves the results. Moreover, we show that valence is better estimated using semantic features while arousal is better estimated using acoustic features.**

I. Introduction

Speech is a natural and effective way of communication between humans, which carries various sources of information about the speaker such as gender, age, physiological and emotional states. Inspired by this, researchers devoted much work into speech analysis, considering it as an efficient method for human computer interaction (HCI) as well. There has been great advances in neutral speech recognition since fifties which could communicate the explicit message given by the speaker. However, humans seldom communicate with neutral speech, mostly there are implicit messages underlying. Emotions are fundamental parts of those implicit messages and for better HCI systems recent research has focused on the recognition of emotional speech. It has been shown that it is crucial for communication systems to recognize humans' emotional states [1].

The task of speech emotion recognition is very challenging and one of the main reasons is the extraction of suitable features [2]. Semantic and acoustic features, which are based on what and how it is said respectively, can be gathered from speech. Most attention has been given to the use of acoustic features for speech emotion recognition [3], [4]. The typical and mostly used features are the pitch, the formants, the short-term energy, the mel-frequency cepstral coefficients (MFCC), and the Teager energy operator based features [5]. Although many acoustic features have been explored, there is no standard group of features defined yet, which efficiently characterize the emotional content, independent of the speaker and the lexical content. Usually many different types of features are combined and feature selection methods are applied to deal with hundreds of features [6]. Once the features are extracted, the choice of a learning model is important as well. Many models have been used for this task, such as hidden Markov model (HMM), support vector machines (SVM) and neural networks, but there is yet no agreement on the most suitable one.

Recently, there has been effort on the integration of acoustic and semantic information of speech [7], [8]. It has been shown in recent studies that combining acoustic and semantic features improve emotion recognition results [8]. Bag-of-words (BOW) features, where each term within a vocabulary is represented by a feature modeled by the term's occurrence frequency within the phrase, and part-of-speech (POS) features, where each phrase is represented by grammatical tags (verbs, nouns, etc.), have been shown to be useful for emotion recognition task [9], [10]. The vocabulary is often limited to predefined emotional keywords including words like 'happy', 'sad' and 'depressed' [8]. However, not only keywords but also general terms may carry the emotional content [11]. The sentences 'I passed the exam' and 'I am happy to have passed the exam' convey similar emotions, but the former does not have the keyword 'happy'. Semantic language analysis, which is the study of 'meaning' and is a subfield of linguistics, could be used as a key to solve this problem. Latent semantic analysis (LSA) is an indexing method that is based on the principle that words occurring in similar contexts are also similar in meanings. It has been used for text-emotion recognition task [12], [13]. It has also been used for the emotional analysis of songs (using lyrics) [14] where they measure the similarities of lyrics as a whole to the emotional keywords defined. The use of LSA can be beneficial also for speech emotion recognition task enabling the use of all the terms in phrases not only the emotional keywords defined, assuming words, or phrases with

similar meanings would carry similar emotions.

Two main and mostly used approaches to emotional modelings are categorical (basic emotions) and dimensional [15]. The categorical approach is based on classification of emotions as basic emotions such as happiness, sadness and surprise that are hard-wired in human brain and recognized universally [16]. On the other hand, within the dimensional approach, researchers argue that emotional states are not independent from one another but related in a systematic manner. To determine the emotional dimensions is a challenge, however, the two dimensions, valence (V) and arousal (A) has been shown to cover the majority of affect variability and been widely used [16], [17]. The arousal dimension represents how excited the emotion is or how much energy is required to express the feeling. Feelings with high arousal induces some physical changes in the body such as increased heart rate, higher blood pressure and greater sub-glottal pressure resulting in change in speech as well such as making it louder, faster and have higher average pitch etc. The valence dimension refers to how positive or negative the emotion is, ranging from pleasant to unpleasant. However, it has not been shown yet how or if the acoustic features correlate with the valence dimension [2]. Even if there exists some works done on dimensional automatic emotion recognition [18], it is still in its infancy [16].

There have been studies in text-based emotion recognition about how to evaluate each word in the emotional dimensions. The affective norms for English words (ANEW) in [19] introduced by Bradley and Lang in 1999, includes a set of normative emotional ratings for 1034 English words. It has been developed to provide researchers with the standardized materials in emotion studies. The self-assessment manikin (SAM), an affective rating systems designed by Lang in 1990 [20], is used to assess the three affect dimensions which are valence, arousal and dominance. The graphic SAM figures has nine values, from low to high and neutral in the middle, comprising bipolar scales in each dimension. ANEW has been widely used for emotion analysis purposes by researchers, yet, how to rate and assess emotions is still questioned.

The acquisition of an appropriate database is another challenge for speech emotion recognition task. While trying to choose a database, there are many points to be considered such as the language, the scope (emotion analysis or recognition), subjects used (adults or children), naturalness (acted or natural), the balance in phrases (the number of phrases per emotion, phrase length, etc.), the emotion model (categorical or dimensional), the assessment type (which emotions for the categorical, continuous or quantized for dimensional) and the duration of phrases or dialogs. Although, there are a number of available databases developed well, it is usually hard to find one that is convenient in all those aspects[1]. Thus, some researchers prefer to design their own emotional speech databases that apply to their research aims [8], [21].

[1]You can check [2] for a review of some commonly used emotional speech databases

In this paper, we design a speech emotion recognition model with a dimensional approach that combines the acoustic and semantic features, taking ANEW words as reference for the emotional keywords. We design our own emotional database using short audio clips from English movies. We make the human subjects rate the emotions expressed in the clips on valence and arousal dimensions. To coincide with the rating procedure of ANEW work, we use the same SAM figures and similar instructions for the subjects. For the acoustic part, using the openEar toolkit [22], we extract and combine hundreds of acoustic features, use the correlation based feature selection (CFS) [23] method to reduce the number of features and use support vector regression (SVR) as the learning model. For the semantic part, we use LSA to find similarities of each term in the text of each clip to ANEW words (word by word similarities) and evaluate the dimensional ratings combining those similarities with A-V values of ANEW words. Finally, we combine acoustic and semantic results and we analyze and discuss the results in both dimensions.

II. Database Design

An emotional speech recognition database that suits the needs of a research project is hard to find especially in the dimensional approach which is not yet commonly studied. We needed a database with English as the language, audio samples consisting of at least a few words to be able to analyze semantically, ratings in valence and arousal dimensions in a way similar to ANEW ratings and adults as the annotators and subjects. Therefore, we designed our own database that consists of short clips taken from English movies.

The clips are rated in valence and arousal dimensions. The response format choice is crucial designing the database since it could affect the resultant ratings. One example is the polarity of the format [24]. Considering a format where the rating 1 represents 'not happy' and 9 'represents' 'happy', if the subjects treat 'not happy' as 'sad', it is bipolar. However, if the aim is a 'unipolar' format where 'not happy' means 'neutral', then this confusion might bring considerable error. To avoid these kinds of error and to coincide with the ANEW work in [19] the results of which are used in the semantic part of our work, we use the same response format and similar instructions they used. The bipolar scale with ratings between 1 and 9 for both dimensions are used with SAM figures [20]. The annotators are asked to rate the emotion expressed in the clips.

The database consists of 59 clips in total from 11 movies with durations between 5 and 25 seconds. The clips include audio and text in the form of subtitles. The audio clips are resampled at 16 kHz. Since the loudness of movies may differ and the loudness is one of the important acoustic features, the long-term loudness[2] of the clips has been normalized using Replay Gain. Replay Gain is an open standard loudness calculation algorithm [25] in which the main idea is to calculate the

[2]Please note that this does not really effect the instantaneous loudness

gain needed on an audio file to match the perceived loudness level of a reference audio file.

A Java applet has been designed for the experiments to get the emotional ratings. In this applet, there are 3 experiments to be carried out by the users and we ask them to fill in some personal information for statistical purposes and a questionnaire at the end to get some feedback. In the first experiment, the clips include just text, in the second just audio and in the third both text and audio. The order of the clips are toggled before each experiment. The experiments take around 1 hour in total. The applet has been made available online [26], thus the users did the experiments wherever they wanted. However, they were asked to do it somewhere not very disturbing and without giving long breaks (the time track of the users have been taken through the applet and been checked). We recruited 13 people (7 female, 6 male), between ages 19 and 28, speaking English fluently. They all claimed that the instructions were clear and they were confident rating.

The Figure 1 shows the arithmetic mean values of the ratings of all clips for the three experiments. The results and more details about the design process can be found in [27].

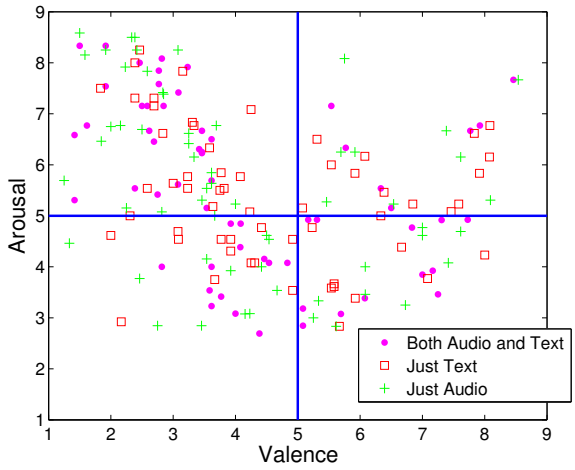

Fig. 1. The mean ratings of all the clips for the three experiments.

III. Modeling Framework

A. An Overview

Figure 2 gives an overview of the modeling modeling framework. The emotions underlying the clips are recognized in arousal-valence (A-V) space in audio and text parts separately and combined in the end to have the final results. The combination is done according to the formula in the equation below,

$$AV_{clip\_comb}(i) = we\_sem(i) * AV_{clip\_sem}(i) + we\_aco(i) * AV_{clip\_aco}(i)$$

where, $AV_{clip\_comb}(i)$, $AV_{clip\_sem}(i)$, $AV_{clip\_aco}(i)$ are the results for the combination, semantic and acoustic parts and $we\_sem(i), we\_aco(i) = (1 - we\_sem(i))$ are the weights of semantic and acoustic parts in the combination respectively for the i^{th} clip.

The details about the audio and text emotion recognition are given in the following sections.

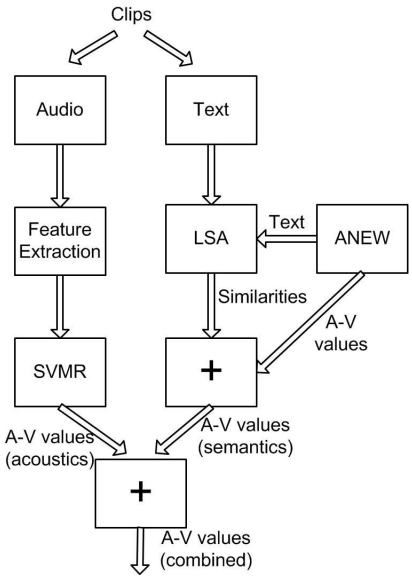

Fig. 2. An overview of the framework. The (+) sign represents that the inputs are combined, the details of the combination methods are in the relative sections.

B. Semantics

We use ANEW words as the reference emotional keywords as mentioned earlier. We use LSA to calculate the word by word similarities between words in the clips and 1034 ANEW words. The following formula can be used to find the estimated A-V values of the corresponding words,

$$AV_{word}(j) = \frac{\sum_{i=1}^{N} AV_{ANEW\_word}(ji) * sim_{ANEW\_word}(ji)}{\sum_{i=1}^{N} sim_{ANEW\_word}(ji)}, \quad (1)$$

where $AV_{ANEW\_word}(ji)$ and $sim_{ANEW\_word}(ji)$ represent the A-V values of the ANEW words and the similarities of the corresponding word (as a result of the LSA algorithm) to ANEW words for the j^{th} clip word to be analyzed and i^{th} ANEW word respectively. N is the number of ANEW words giving similarities more than a threshold value of sim_thr. Thus, N words above the threshold are taken into account within this formula. However, in the case of similarity score of 1 to one of the words, $(N-1)$ words with similarities above the threshold would still be taken into account which we would like to avoid. The increased threshold could be a solution to this case, yet, it could affect badly the cases in which all the similarities are lower than that threshold. Thus we define a weight for each word using the following formula,

$$we_{ANEW\_word}(ji) = sim_{ANEW\_word}(ji)^{\frac{we\_thr}{1+gamma-max\_sim}} \quad (2)$$

where $we_{ANEW\_word}(ji)$, we_thr and max_sim represent the weight for each ANEW word for the j^{th} clip word, a threshold value to be optimized between 0 and 1 and the maximum similarity of j^{th} clip word to ANEW words respectively. $gamma$ is a very small number (close to zero) used to avoid dividing by zero in the case of max_sim of 1. Figure 3 illustrates the weight-similarity relation visually. Then, we insert $we_{ANEW\_word}(ji)$ instead of $sim_{ANEW\_word}(ji)$ in equation 1 to estimate AV values of a word in a clip.

Finally, A-V values of a clip are estimated using the estimated A-V values of each word in it in an equation similar to 1. The weight of each word's contribution is taken as the maximum similarity found for that word using LSA and then the results is normalized by dividing by the sum of the weights of each word comprising the clip.

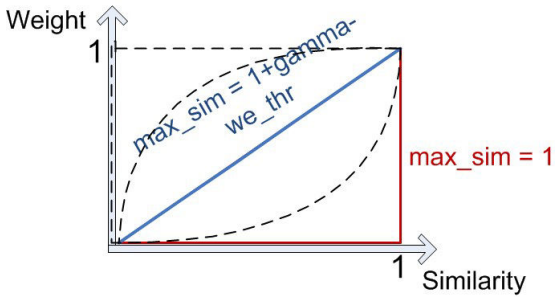

Fig. 3. Similarity to weight conversion using Equation 2. The curve shape changes with we_thr and max_sim. The red and blue colored curves show two different specific cases.

C. Acoustics

There is no group of features yet which efficiently characterize the emotional content, independent of the speaker and the lexical content as stated before. This is why many researchers prefer to combine many acoustics features such as the pitch, the formants, MFCC and energy and then apply a feature selection algorithm to find best features for the current work. We use the openEar toolkit [22] which is an open-source affect and emotion recognition engine. It does not only enable us to extract many acoustic features, but also to apply a feature selection algorithm and a recognition engine.

OpenEar provides low-level audio features such as formants, pitch, MFCC, linear predictive coding coefficients (LPCC), zero-crossing rate, signal energy and voice quality and features obtained by applying various statistical functionals and transformation to those features. We have 988 features in total the details of which can be found in [22]. As a feature selection method we use the Correlation Feature Selection (CFS) [23] which is based on the hypothesis that good features are the ones which do not correlate with each other but correlate with the classification. Finally, as the recognition engine we have SVR which is a maximum margin algorithm, computing a linear regression function in a high dimensional kernel induced feature space where the input data is using a nonlinear transformation [28].

IV. Experiments

The outliers of the database are detected using Peirce's criterion [29] and rejected. An LSA software package [30], that is based on term frequency inverse document frequency (TF-IDF) weighting and that outputs cosine distance as the similarity measure, has been used. The corpus used for LSA is called HAWIK combining Harvard Classics literature samples with Wikipedia articles and Reuters news items. HAWIK corpus has been used in [14] and has been shown to perform well for affect recognition purposes.

The database we designed has been divided into a train set with 29 clips and a test set with 30 clips. The optimization of the parameters for the text part and the combination part has been done using the train set, and the final results are evaluated using the test set. For the acoustic part, all 59 clips have been used to find the optimal parameters using 5-fold cross validation method of the openEar toolkit. The final results are given using the leave-one-out method.

We measure the error between the estimated and human-rated A-V values using mean absolute error (MAE) which gives the average of the absolute differences between them. We also check the root mean squared error (RMSE) for the final results which is a useful error measure for the applications where large errors might be specifically undesirable.[3]

We looked for the optimal weights for semantic and acoustic parts (we_sem and we_aco as in Equation 2) for the combination of the results of the two giving minimum MAE and RMSE using the train set. Figure 4 gives the results using we_sem values between 0 and 1. Thus, we choose we_sem giving minimum MAE and RMSE to be 0.8 and 0.85 for the valence dimension, and 0 and 0.2 for the arousal dimension respectively (then, we_acc is simply equal to (1-we_sem)). All the optimal parameters are found similarly minimizing MAE and RMSE. For the text part, sim_thr and we_thr are 0.2 and 0.3 respectively.

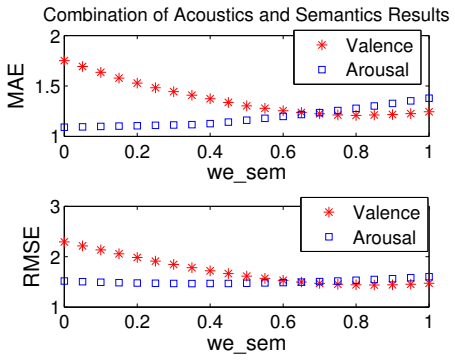

Fig. 4. Mean Absolute error (MAE) and Root Mean Squared Error (RMSE) versus the weight of semantics (we_sem) in the combination, using the train set.

We also look for a subset of ANEW words that could be better at helping recognizing affect. Although, it is convenient

[3]The errors are squared before they are averaged, so higher weight is given for large differences.

and useful to use ANEW words since they are already rated emotionally, we do not know if the choice of those specific words are the optimum for our work. We created a subset of ANEW words (116 words), which are affect related like adjectives 'afraid', 'depressed' or nouns like 'fear' and 'hate'. Figure 5 gives the MAE results (RMSE results are similar) of the train set using the subset and all ANEW words. We use the subset to obtain the final results.

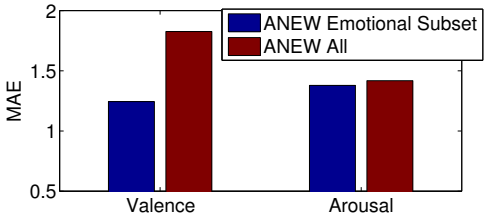

Fig. 5. Mean Absolute error (MAE) results for both dimensions using whole ANEW words or the emotional words subset, using the train set.

For acoustic features, the frame size is set to 25 ms at a skip rate of 10 ms. The frequency range of the spectrum is set from 0 to 8 kHz. After applying CFS, around 40 and 70 features are selected for the valence and arousal dimensions respectively (the difference is due to selected loudness related features for the arousal). We use a radial basis function (RBF) type kernel for the valence dimension. The linear type kernel is used for the arousal dimension, since the number of data samples is lower than features. THe details about all the parameters for the acoustics part can be found in [22].

V. Results and discussion

Table I gives the results for the semantic and acoustic parts using the test set with optimal parameter values (optimization process is explained in section IV). It is hard to compare our results and make a strong conclusion about recognition performance, since the dimensional affect recognition research area is quite recent. Moreover, it is also hard to compare them with limited number of previous works, since the databases used are different. However, in [22], introducing the openEar toolkit, they report MAE of 0.28 and 0.38 for A-V dimensions respectively for the SAL corpus [22], which is a continuous dimensional affect corpus of natural speech for A-V values rated between -1 and 1. If we normalize their result to match ours, using the fact that A-V rated values for our database is in the range of 1 to 9, their results can be considered as 1.12 and 1.52 for A-V dimensions which is comparable to our results.

Although, we are mostly satisfied with recognition results in semantic and acoustic parts, we are mainly interested in the results obtained by combining them to see if we have the desirable and expected improvement and the effect of it in the two dimensions. For the weights of semantics and acoustics in A-V dimensions, we use the optimal values found as described in section IV and as shown in Figure 4. We observe, as seen in Table II that the results are slightly better than the best of semantics and acoustics results in both

		Semantics	Acoustics
Valence	MAE	1.45	1.98
	RMSE	1.85	2.50
Arousal	MAE	1.39	1.29
	RMSE	1.64	1.54

TABLE I
Mean Absolute error (MAE) and Root Mean Squared Error (RMSE) for semantic and acoustic parts in A-V dimensions, using the test set

dimensions. Thus, our results show that combining acoustic and semantic information improves the recognition results also for continuous dimensional approach as it was shown before for the categorical approach in previous works [8].

The most interesting and a novel result in our work is that we show that different weights for semantic and acoustic parts for the combination are needed, which can be seen clearly on Table II. We show that the valence dimension is recognized better using semantic features while the arousal dimension using acoustic features. We interpret this result as, *the valence dimension is more about what we say, while the arousal dimension is more about how we say it*. This result agrees with the fact that it has not been shown yet how or if the acoustics features correlate with the valence dimension [2] as stated before. Moreover, our results also coincides with the results in [31] where they design an emotion recognition model in A-V space as well using bio-sensors extracting features such as body temperature, breath speed and heart activation. Their results show that using bio-sensors, it is much harder to estimate the valence dimension than the arousal. In other words, what they show is that physiological changes in the body gives more information about the arousal dimension. Therefore, since, the physiological changes affect our speech as well, specifically how we speak, our interpretation that 'the arousal dimension is more about how we say it' is supported by their results.

		Weights (we_sem / we_ac)	Combined Result
Valence	MAE	0.80 / 0.20	1.40
	RMSE	0.85 / 0.15	1.77
Arousal	MAE	0 / 1	1.28
	RMSE	0.20 / 0.80	1.52

TABLE II
Mean Absolute error (MAE) and Root Mean Squared Error (RMSE) for the combination of semantics and acoustics parts in A-V dimensions with weight values of each, using the test set

VI. Conclusion

This paper described a method of recognizing affect in speech with a dimensional approach and combining semantic and acoustic features. We created a corpus that consists of short movie clips that contain audio and text in a subtitle

format, rated in arousal and valence dimensions by human subjects. We showed that combining the semantic and acoustic features improve the recognition results. We also showed that, semantic features are better at estimating the valence dimension while the acoustic features are better at eliciting the arousal dimension.

VII. Acknowledgments

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886 and in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. The publication only reflects the authors' views.

References

[1] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[2] M. El Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] T. Nwe, S. Foo, and L. De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[4] D. Ververidis and C. Kotropoulos, "Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm," in *IEEE International Conference on Multimedia and Expo, 2005. ICME 2005*. IEEE, 2005, pp. 1500–1503.

[5] Ververidis, D. and Kotropoulos, C., "Emotional speech recognition, resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[6] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *IEEE International Conference on Multimedia and Expo, 2005. ICME 2005*. IEEE, 2005, pp. 864–867.

[7] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04)*, vol. 1. IEEE, 2004, pp. I–577.

[8] Z. Chuang and C. Wu, "Multi-modal emotion recognition from speech and text," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 9, no. 2, pp. 45–62, 2004.

[9] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous *et al.*, "Whodunnit-searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, 2011.

[10] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "Combining efforts for improving automatic classification of emotional user states," *Proc. IS-LTC*, pp. 240–245, 2006.

[11] C. Wu, Z. Chuang, and Y. Lin, "Emotion recognition from text using semantic labels and separable mixture models," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 5, no. 2, pp. 165–183, 2006.

[12] A. Gill, R. French, D. Gergle, and J. Oberlander, "The language of emotion in short blog texts," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 2008, pp. 299–302.

[13] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*. ACM, 2008, pp. 1556–1560.

[14] M. Petersen and L. Hansen, "Modeling lyrics as emotional semantics," *Proceedings of YoungCT, KAIST Korea Advanced Institute of Science and Technology*, 2010.

[15] D. Grandjean, D. Sander, and K. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and cognition*, vol. 17, no. 2, pp. 484–495, 2008.

[16] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[17] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[18] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective mtv analysis based on arousal and valence features," in *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 2008, pp. 1369–1372.

[19] M. Bradley and P. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," *University of Florida: The Center for Research in Psychophysiology*, 1999.

[20] Bradley, M.M. and Lang, P.J., "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[21] Y. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *Proceedings of International Conference on Machine Learning and Cybernetics*, vol. 8. IEEE, 2005, pp. 4898–4901.

[22] F. Eyben, M. Wollmer, and B. Schuller, "Openear introducing the munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–6.

[23] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[24] J. Russell and J. Carroll, "On the bipolarity of positive and negative affect." *Psychological Bulletin*, vol. 125, no. 1, pp. 3–30, 1999.

[25] D. Robinson, "Replay gain," 2001, `http://www.replaygain.org/`.

[26] S. G. Karadogan, "Emotional movie database (emov) creation," 2011, `http://www.student.dtu.dk/~seka/`.

[27] Karadogan, S. G., "Emotional speech database design details and the java applet code written for the design," 2012, `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6231`.

[28] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[29] S. Ross, "Peirce's criterion for the elimination of suspect experimental data," *Journal of Engineering Technology*, vol. 20, no. 2, pp. 38–41, 2003.

[30] DTU, "Latent semantic analysis software application written in java including hawik and lywik corpora," 2010, `http://dl.dropbox.com/u/5442905/lsa.zip`.

[31] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," *Affective Dialogue Systems*, pp. 36–48, 2004.