Annotation and classification of political advertisements

Samuel Kim[1,2]*, Panayiotis Georgiou*[2] *and Shrikanth Narayanan*[2]

[1] DSP Lab., Yonsei University, Seoul, Korea
[2] SAIL Lab., University of Southern California, Los Angeles, USA

samuel.kim@dsp.yonsei.ac.kr

Abstract

Political advertising has changed drastically over the last several decades with video advertising becoming a major force in all outlets from the traditional TV medium to online media. In this work, we attempt to automatically classify the political advertisements along various dimensions such as `purpose`, `content` and `emotion`. First, we use a crowd-sourcing method to annotate the political advertisements in terms of how viewers perceive them along the above mentioned aspects. Then, we use audio-based features and machine learning algorithms for automatic classification tasks. In particular, we deploy speech-related features along with support vector machine (SVM) and music-related features along with k-nearest neighbor (KNN). The analysis of crowd-sourced annotations shows that the same advertisements are often used to serve multiple `purpose` and that certain `content` categories such as speech clips from the candidate and other public figures are more prevalent. The experimental results using speech/audio features on advertisements aired during the U.S. presidential campaign of 2012 show promising classification performance.

1. Introduction

Video advertising, such as those on television or online media, is one of the most effective ways to advertise political issues to the voters; usually these are in a format where a politician or a party presents their position and, increasingly, the advertisements include content related to the topic, either a positive portrayal of the candidate or negative for the opponent, without the candidates themselves directly appearing. Most advertisements are created for TV and thus are short in duration (∼30 sec.) although recently we are seeing expanded versions of these appearing online. The goal of this feasibility study is to determine whether conventional content-based approaches using speech and audio information can offer meaningful analytics on political advertisements. Particularly, we attempt to automatically classify the political advertisements with respect to characteristics that are of relevance to human judgement. Another motivation for this work comes from social and political science studies which have extensively investigated how political advertisements affect the electorate according to the intended goals and contents of the advertisements: negative/positive advertisements [1, 2, 3] and image/issue advertisements [4, 5].

Studies on political audio-visual media from a computational content processing angle are not new. Several studies in the literature address the domain of political debates on television where a small group of politicians debate, usually moderated by journalists, over a relatively long period (30∼100 minutes). Gregory and Gallagher used candidates' nonverbal vocal features to predict the outcome of the 19 televised U.S. presidential debates of eight elections between 1960 and 2000 [6]. They found candidates' non-verbal vocalizations provide a precise metric of the dominance or commanding presence, which are in turn highly correlated with the popular vote percentages. Kim *et al.* used conversational patterns and vocal prosody statistics to detect conflicts during political debates [7]. Using 45 political debates broadcast in Switzerland, they showed promising results in detecting conflicts [7] and conflict escalations [8]. More recently, Kaplan and Rosenberg tried to predict the winner of the presidential election using language transcript based features extracted from 25 U.S. presidential debates between 1976 and 2008 [9]. Their approach can predict the winners of elections moderately above chance level.

The focus of this work is towards retrieval of ads based on qualitative (human-centered) dimensions and we address that through two main contributions. First we introduce a crowd-sourcing method to annotate political advertisements in terms of how viewers judge them along various qualitative dimensions, such as the `purpose`, `content` and `emotion`. We designed a questionnaire to easily annotate the advertisements that are typically heterogeneous in terms of their intended goals as well as content. For example, the same advertisement can be produced to promote a candidate (positive messaging) and to attack the opponent at the same time (sending a negative message). Furthermore, in terms of content, an advertisement can contain multiple content such as professional narration, testimony from individuals, and candidate's own public speech excerpts. The second contribution of this paper is to perform automatic classification of the political ads with respect to the above qualitative dimensions using audio signal based features and machine learning algorithms. The rationale behind this is that audible information embedded in the advertisements is highly correlated with the target perception of the viewers.

2. Annotation

2.1. Database

We constructed our corpus with the help of Francis Steen (UCLA) and Mark Turner (Northwestern) and through the *NewsScape Library of International Television News* [10, 11]. The advertisements used for this study are from the 2012 U.S. presidential campaign between Barack Obama and Mitt Romney. We selected the ads that were under 60s in length (the longest TV ads of the campaign) thus focusing on broadcast

Table 1: Statistics of database

Candidate	Obama	Romney	Total
Number of clips	172	104	276
Average length of clips	35.6	35.4	35.5

The aurthors like to thank Profs. Francis Steen (UCLA) and Mark Turner (Northwestern) for providing access to the *NewsScape Library of International Television News*.

media [12]. As shown in Table 1, the collected political advertisements include 276 clips (172 for Obama and 104 for Romney) whose average length is 35.5 seconds (total 2.7 hours). The clips are stored in H.264 codec (30 frames per second) and MPEG-4 AAC codec (44.1 kHz stereo) format. To extract audio-related features, we down-sampled and merged the audio signals into 16 kHz mono using the FFmpeg toolkit (http://ffmpeg.org/).

2.2. Crowd-sourced Annotation Process

We used a crowd-sourcing strategy to annotate the entire dataset. Specifically, we used *Amazon Mechanical Turk* (MTurk, https://www.mturk.com/) to easily manage the crowd-based annotation process [13]. We prepared a questionnaire that consisted of 5 different questions to label the political advertisements along different dimensions of interest namely, `purpose`, `content` and `emotion`, as shown in Table 2. After an annotator had watched through a given advertisement video clip, he/she was asked to answer the questionnaire.

The first question was designed to filter out possible outlier annotators; we compare the annotators' answers and the ground truth of advertisements (i.e., which candidate made the advertisement). The questions considering the `purpose` and the `content` of the advertisements are designed to allow multiple `purpose` and `content` labels so that they can be banalized, i.e., 0 or 1. Another set of questions was given to the annotators to estimate judgement of expressed emotions [14]. In particular, we use emotion primitive descriptors, i.e., *valence* and *activation*, and the annotators were asked to select one answer out of five possible alternatives $\{1, 2, 3, 4, 5\}$ which are in an ordinal scale accompanied by emoticon images, a.k.a., self assessment manikins (SAM).

The answers are averaged across the annotators and split into binary classes; the class of i-th advertisement for q-th question is determined as follows.

$$c_{i,q} = \begin{cases} 1 & s_{i,q} \geq \theta_q; \\ -1 & \text{otherwise}, \end{cases}$$

where θ_q is a tunable threshold. In this work, we use $\theta_q = 0.5$ for the sub-questions considering `purpose` and `content` to represent that more than half annotators agree on the existence of such purpose or content. We also set $\theta_q = median(s_{i,q})$ for the questions considering emotion primitive descriptors.

2.3. Annotation Analysis

First, we analyze the sub-questions related to the perceived `purpose`. Table 3 shows that the number of clips that are labeled with respect to the target `purpose` and the percentage of the corresponding clips that are co-labeled as another `purpose`. For example, there are 178 clips that are labeled as PR (promoting candidate) and 27% of them are also labeled as AT (attacking opponent). This illustrates that the political advertisements are often designed to promote their candidate and attack the opponent at the same time. Note that RE (responding to attacks) is hardly observed in the database we used.

Table 4 shows the number of clips that are labeled with respect to the `content` and the percentage of the corresponding clips that are labeled with respect to the `purpose`. For example, there are 74 clips that are labeled as PS (public speech) and 66.2% of the clips are labeled as PR and 67.6% are labeled as AT. Note that all the clips are labeled as DS (designated speech) are also labeled as PR. This just implies that candidates tend to promote themselves when they appear on the advertisement speaking to the viewers and it is consistent with the findings in [5]. It is also noticeable that other public figures and supporters' testimonials more frequently used to promote the candidate than to attack the opponent. On the other hand, professional narrations are often used to attack the opponent while music and candidates' public speech can be equally used for both supporting and attacking a candidate. The analysis results are intuitively reasonable and partially supported by political science research; the candidate himself and other public figures tend to be careful about directly attacking the opponent since it may cause negative effects on themselves [1, 2]. Instead, when the candidate wants to attack their opponent in an advertisement because *"no one likes them, but they work"* [2], they use narrators to attack the opponent without tarnishing their image directly.

Table 5 shows the correlation values between the scores for the `purpose` and the `emotion` primitives. As shown in the table, valence is positively correlated with PR and negatively correlated with AT. This trend is reasonable in the sense that attacking the opponent might include negative aspects of the opponent, while promoting the candidate should elaborate the positive aspect of the candidate. On the other hand, there are only relatively loose correlations between activation and the scores of PR or AT indicating an even keel in the content expression. Similar trends can be observed in Fig. 1 which depicts the distribution of scores according to the labeled `purpose`.

Table 2: Annotation questionnaire and possible answers.

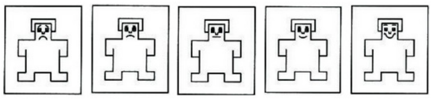
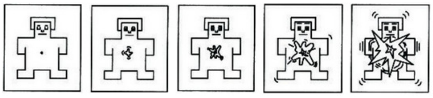
Question	[Abbr.]
Which candidate made this advertisement?	
□ Barack Obama	
□ Mitt Romney	
What is the `purpose` of this advertisement?	
□ Promoting the candidate	[PR]
□ Attacking the opponent	[AT]
□ Responding to attacks	[RE]
What is the `content` of this advertisement?	
□ Candidates' public speech	[PS]
□ Candidates' speech designated for this ad	[DS]
□ Other public figures	[PF]
□ Professional Narrator	[PN]
□ Testimonials	[TM]
□ Music and text	[MU]
How positive/negative is the advertisement?	[Valence]
	
How passive/active is the advertisement?	[Activation]
	

Table 3: Number of clips that are labeled with respect to `purpose` and the percentage of the corresponding clips that are co-labeled with another `purpose`.

`Content`	# of clips	Percentage of purpose (%)		
		PR	AT	RE
PR	178	-	27.0	1.7
AT	142	33.8	-	2.8
RE	9	33.3	44.4	-

Table 4: Number of clips that are labeled with respect to content and the percentage of the corresponding clips that are labeled with respect to purpose.

Content	# of clips	Percentage of purpose (%)		
		PR	AT	RE
PS	74	66.2	67.6	2.7
DS	22	100.0	40.9	4.5
PF	50	74.0	26.0	6.0
PN	130	52.3	71.5	3.8
TM	50	76.0	32.0	0.0
MU	198	58.6	63.1	3.0

Table 5: Correlation between the scores for purpose and scores for emotion primitives ($^{*}p < 0.01, ^{**}p < 10^{-50}$).

	PR	AT	RE
Valence	0.85 **	– 0.93 **	0.04
Activation	– 0.27 *	0.19 *	0.08

Table 6: Correlation between features and individual questions. Three most correlated features (in absolute values) are selected.

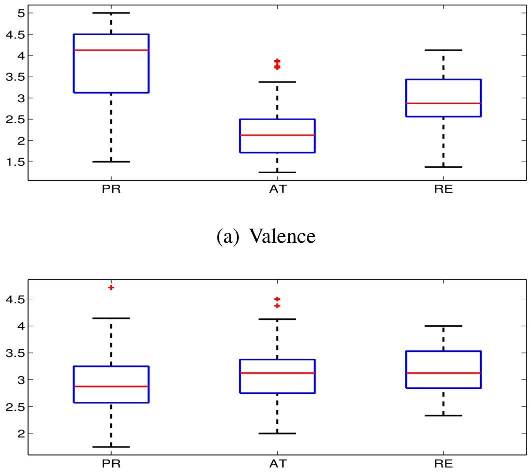
Label	Feature name	cc
PR	mfcc_sma[1]_pctlrange0-1	-0.45
	mfcc_sma[1]_percentile99.0	-0.42
	mfcc_sma[1]_stddev	-0.42
AT	mfcc_sma[1]_stddev	0.45
	mfcc_sma[1]_pctlrange0-1	0.45
	mfcc_sma[1]_linregerrQ	0.45
RE	lspFreq_sma_de[5]_linregc1	0.26
	lspFreq_sma_de[4]_linregc1	0.25
	pcm_loudness_sma_de_linregc1	-0.24
PS	mfcc_sma_de[9]_kurtosis	0.49
	F0finEnv_sma_quartile3	0.42
	lspFreq_sma[0]_quartile2	0.42
DS	lspFreq_sma[4]_percentile99.0	0.22
	lspFreq_sma_de[6]_skewness	0.21
	jitterLocal_sma_de_kurtosis	0.21
PF	logMelFreqBand_sma[1]_iqr2-3	0.34
	logMelFreqBand_sma[2]_iqr2-3	0.32
	logMelFreqBand_sma_de[2]_quartile2	-0.31
PN	mfcc_sma_de[1]_stddev	0.66
	mfcc_sma_de[1]_linregerrQ	0.65
	logMelFreqBand_sma_de[7]_percentile1.0	-0.65
TM	lspFreq_sma[4]_upleveltime90	-0.27
	logMelFreqBand_sma[7]_linregerrA	-0.26
	lspFreq_sma_de[2]_minPos	0.25
MU	logMelFreqBand_sma[1]_iqr1-3	-0.41
	logMelFreqBand_sma[0]_quartile1	0.40
	logMelFreqBand_sma[2]_iqr1-3	-0.40
Valence	mfcc_sma[1]_stddev	-0.47
	mfcc_sma[1]_pctlrange0-1	-0.47
	mfcc_sma[1]_linregerrQ	-0.47
Activation	logMelFreqBand_sma[1]_percentile99.0	0.32
	logMelFreqBand_sma[4]_quartile3	0.31
	logMelFreqBand_sma[2]_quartile3	0.30

3. EXPERIMENTS ON AUTOMATIC Classification

The goal of this section is to explore if it is possible to automatically classify the above mentioned human-derived descriptors with features extracted from audio signals. We consider individual labels separately and perform classification tasks independently while using the same feature sets.

3.1. Speech-related features

We use the open-source Emotion and Affect Recognition Toolkit's feature extracting algorithm, i.e., openSMILE [15]. Particularly, we employ the same setting as in the INTERSPEECH Paralinguistic Challenge 2010 [16] which extracts 1582 acoustic features. The features include various functionals such as mean, standard deviation, quantiles, etc. of 38 low-level descriptors such as F0, MFCC, loudness, etc. This set of features has been shown to be promising in various paralinguistic classification tasks such as recognizing age, gender and emotion of speakers (see [16] for more details).



(a) Valence

(b) Activation

Figure 1: Score distributions of (a) valence and (b) activation according to the labeled purpose.

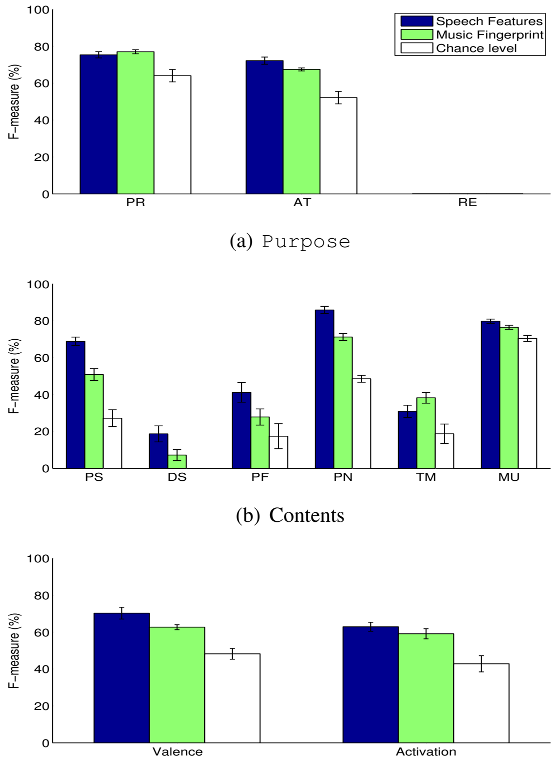
Table 6 shows the correlation values between the features and the scores of individual questions. The three most correlated features, either negatively or positively, are selected in the list. As shown in the table, functionals of MFCC and logMelFreqBand often appear as one of the most meaningful features. Notably, they dominate the top three features for PR, AT, PF, PN, MU, Valence and Activation. This indicates that the spectral characteristics of audio signals, especially captured in MFCC and log-energies in mel-frequency bands, are important for studying the labels of interest.

For classification tasks, we use a simple linear-kernel based SVM using the LIBSVM toolkit [17].

3.2. Music Fingerprint using Chroma-based features

Advertisements often include music, both in the background and foreground. We adopt the music fingerprint extraction method proposed in [18, 19] to measure similarities between two different pieces of music. The method uses the chroma feature which represents the energy distribution over the twelve Western chromatic pitch classes (A to G#) and computes a covariance matrix of the chroma feature vectors as a music fingerprint, i.e., $\mathbf{\Phi} = E\left[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T\right]$ where $\mathbf{x}$ represents the twelve-dimensional chroma feature vector and T represents the matrix transpose. The music fingerprint provides the relationship between individual pitch classes, which leads to modeling the harmony structure of a given music piece.

Then a simple template matching method is used to measure the similarity of the two music fingerprints. The similarity between music i and j is computed as follows; $\sigma_{ij} = \|\mathbf{\Phi}_i * \mathbf{\Phi}_j\| / \|\mathbf{\Phi}_i\| \, \|\mathbf{\Phi}_j\|$ where $\|\mathbf{\Phi}\| = \sqrt{\sum_k \sum_l (\phi_{kl})^2}$ and ϕ_{kl} represents the k-th row and l-th row element of the music fingerprint $\mathbf{\Phi}$; the $*$ operator represents a pairwise multiplication of each element of matrix. It should be noted that even in playing the same music, it is possible to transpose the key of the music. To compensate for the possible key transposition, we



(a) `Purpose`

(b) Contents

(c) Emotion

Figure 2: Performance in classifying various aspects of political ads using audio.

circularly shift one of the fingerprints in the diagonal direction by one semitone step and take the maximum similarity value.

For classification tasks, we use a simple KNN method utilizing the similarity measure. For simplicity, we set $k = 1$ which will yield the class of the closest training sample as the result of classification task.

3.3. Experimental Results

Experiments are performed using a 5-fold cross validation; the entire dataset is randomly split into 5 folds where 4 are used as training and the remaining is used for testing. The procedure is repeated until all the folds are used for testing. We run this 5-fold cross-validation 10 times and average the performance. For comparison, we also measure chance level classification performance by assigning random labels based on prior probability of labels by counting the number of instances in training folds. The performance of the classification tasks are given in terms of F-measure which is the harmonic mean of precision and recall.

Figs. 2 (a), (b) and (c) show the performance of the classification tasks in terms of F-measure according to different types of classes for `purpose`, `content` and `emotion` primitives, respectively. Note that all differences between different methods, i.e., speech features, music fingerprint and chance level, within a class are significant (t-test with $\alpha = 0.05$) except for the difference between music fingerprint and chance level in detecting PF. In detecting `purpose`, as shown in Fig. 2 (a), detecting whether the advertisement includes promotion or attack of a candidate can be achieved with 75.4±1.7% and 72.2±2.0% with the speech features and 77.4±1.1% and 67.5±0.8% with the music fingerprint approach. However, the system fails to detect the advertisements that are designed for responding to attacks. This may be because there are too few samples, i.e., 9 instances, or that the employed feature vectors are not able to capture the characteristics of ads responding to attacks. As shown in Fig. 2 (b), the performance varies according to the `content` label. For example, detecting narration and music performs relatively well (85.9±2.0% and 79.8±1.1% with the speech features and 71.2±1.9% and 75.8±1.1% with the music fingerprint approach) while detecting speech designated for the ads performs relatively poor (18.7±4.3% with the speech features and 7.2±3.0% with the music fingerprint approach). The classification tasks using the speech-related features usually outperform the music-related features except in detecting testimonials. Note that even in detecting music content speech-related features outperform the music-related features (the usefulness of spectral envelope features such as MFCCs for music processing in fact is well known). It may be because the goal of the music fingerprint method is to measure the similarity of harmony structure regardless of existence of musical content.

The results for detecting emotion primitives also seem promising and significantly outperform chance level, as shown in Fig. 2 (c). The F-measures in detecting valence and activation are 70.3±3.2% and 62.9±2.5% with the speech features and 62.7±1.4% and 59.1±2.7% with the music fingerprint approach. It is interesting that the speech-related features that are extracted by openSMILE [15] work well even with highly heterogeneous sound media of advertisements. The performance with the music fingerprint approach implies that similar music content, especially similar harmonic structures, are used in ads of similar emotional content.

4. Conclusions and Future Work

We analyzed political advertisements of the U.S. presidential campaign 2012. Video advertising is considered as one of the most effective ways in announcing political issues and many social and political scientists have been studying the content of the advertisements targeting the electorate, and their effects, especially those of negative/positive advertisements.

In this work, we studied the political advertisements along various dimensions, such as `purpose`, `content`, and `emotion` of the ad. We introduced a crowd-sourcing method to annotate the audio-visual political advertisements in terms of how the viewers judge them. We performed automatic classification with respect to `purpose`, `content`, and `emotion` using features extracted from the audio signals (speech-related features and music-related features) and machine learning algorithms (SVM and KNN). The experimental results show that detecting the intended goals of the advertisements, i.e., if the advertisements are designed to promote the candidate or to attack the opponent, and detecting the expressed emotions are feasible with audio-based features. Detecting the `content` of the advertisements varies depending on the specific content category. Although is aimed at application such as content-based information retrieval, some of the findings are consistent with existing work in social and political studies.

While this paper treats individual descriptors independently, in the future, we will exploit co-occurrence of `purpose`, `content`, and `emotion` labels within the classification framework. We also plan to analyze visual features with respect to the labels used in this work.

5. References

[1] W. Schenck-Hamlin, D. Procter, and D. Rumsey, "The influence of negative advertising frames on political cynicism and politician accountability," *Human Communication Research*, vol. 26, no. 1, pp. 53–74, Jan. 2000.

[2] R. R. Lau, L. Sigelman, C. Heldman, P. Babbitt, L. E. E. Sigelman, and G. Washington, "The Effects of Negative Political Advertisements : A Meta-Analytic Assessment," *American Political Science Association*, vol. 93, no. 4, pp. 851–875, 1999.

[3] L. L. Kaid and A. Johnston, "Negative versus positive television advertising in U.S. presidential campaigns, 1960–1988," *Journal of Communication*, vol. 41, no. 3, pp. 53–064, 1991.

[4] E. Thorson, W. G. Christ, and C. Caywood, "Effects of issueimage strategies, attack and support appeals, music, and visual content in political commercials," *Journal of Broadcasting & Electronic Media*, vol. 35, no. 4, pp. 465–486, 1991.

[5] A. Johnston and L. L. Kaid, "Image Ads and Issue Ads in U.S. Presidential Advertising: Using Videostyle to Explore Stylistic Differences in Televised Political Ads From 1952 to 2000," *Journal of Communication*, vol. 52, no. 2, pp. 281–300, Jun. 2002. [Online]. Available: http://doi.wiley.com/10.1111/j.1460-2466.2002.tb02545.x

[6] S. W. Gregory and T. J. Gallagher, "Spectral analysis of candidates ' nonverbal vocal analysis spectral communication : U.S. presidential election predicting outcomes," *Social Psychology Quarterly*, vol. 65, no. 3, pp. 298–308, 2002.

[7] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: ratings and analysis of broadcast political debates," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2012, pp. 5089–5092.

[8] S. Kim, S. H. Yella, and F. Valente, "Automatic detection of conflict escalation in spoken conversations," in *Proceedings of INTERSPEECH*, Sep. 2012.

[9] I. Kaplan and A. Rosenberg, "Analysis of speech transcripts to predict winners of U.S. presidential and vice-presidential debates," in *IEEE Workshop on Spoken Language Technology*, 2012, pp. 449–454.

[10] M. B. Turner and F. F. Steen, "Multimodal construction grammar," in *Available at SSRN: http://ssrn.com/abstract=2168035 or http://dx.doi.org/10.2139/ssrn.2168035*, October 29, 2012.

[11] The red hen website. [Online]. Available: https://sites.google.com/site/distributedlittleredhen

[12] P. Rosenthal and F. Steen, "UCLA communication studies news archive," *URL: http://www.sscnet.ucla.edu/tna/setesting*, 2006.

[13] C. Callison-Burch and M. Dredze, "Creating speech and language data with Amazon's Mechanical Turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, ser. CSLDAMT '10. Association for Computational Linguistics, 2010, pp. 1–12.

[14] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimations of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462. [Online]. Available: http://doi.acm.org/10.1145/1873951.1874246

[16] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *In Proceedings of INTERSPEECH*, Makuhari, Japan, Sep. 2010.

[17] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/$\ sim$cjlin/libsvm.

[18] S. Kim and S. Narayanan, "Dynamic chroma feature vectors with applications to cover song identification," in *IEEE International Workshop on Multimedia Signal Processing*, 2008, pp. 984–987.

[19] S. Kim, E. Unal, and S. Narayanan, "Music fingerprint extraction for classical music cover song identification," in *International Conference of Multimedia and Expo*, 2008, pp. 1261–1264.