Prosodic cues for emotion: analysis with discrete characterization of intonation

Houwei Cao[1], *Štefan Beňuš*[2,3], *Ruben C. Gur*[1], *Ragini Verma*[1] *and Ani Nenkova*[1]

[1] University of Pennsylvania, United States
[2] Constantine the Philosopher University, Nitra, Slovakia
[3] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia

Houwei.Cao@uphs.upenn.edu, sb513@nyu.edu, gur@mail.med.upenn.edu,
Ragini.Verma@uphs.upenn.edu, nenkova@seas.upenn.edu

Abstract

In this paper we study the relationship between acted perceptually unambiguous emotion and prosody. Unlike most contemporary approaches which base the analysis of emotion in voice solely on continuous features extracted automatically from the acoustic signal, we analyze the predictive power of discrete characterizations of intonations in the ToBI framework. The goal of our work is to test if particular discrete prosodic events provide significant discriminative power for emotion recognition. Our experiments provide strong evidence that patterns in breaks, boundary tones and type of pitch accent are highly informative of the emotional content of speech. We also present results from automatic prediction of emotion based on ToBI-derived features and compare their prediction power with state-of-the-art bag-of-frame acoustic features. Our results indicate their similar performance in the sentence-dependent emotion prediction tasks, while acoustic features are more robust for the sentence-independent tasks. Finally, we combine ToBI features and acoustic features together and further achieve modest improvements in sentence-independent emotion prediction, particularly in differentiating fear and neutral from other emotion.
Index Terms: ToBI, emotion, automatic prediction

1. Introduction

Despite clearly perceived connection between emotional meanings and prosody, a satisfactory model linking the two has been elusive. Both prosody and emotion are complex phenomena and the adopted framework for their representation vastly influences the interpretability of discovered relationships between the two. The main question for prosody analysis in emotional speech research is how to represent the melody contour. Most research relies on continuous features extractable automatically from the acoustic signal for studying the relationship between emotions and prosody [10]. Discrete representations of prosody in the ToBI [20] or Tilt representations [21] on the other hand mark perceptually salient properties of the utterance tune.

Given that the nature of the precise mapping between the underlying phonological representation of intonation and its phonetic implementation is not known, such labeling might tap into a different type of information regarding the relationship between emotion and prosody. Mozziconacci [14] reviewed several studies including Scherer et al. [19] and Mozziconacci [13] arguing that both the representations of F0 contour using models traditionally considered phonological as well as those representing phonetic implementation of pitch in terms of levels and ranges, offer independent, and possibly additive information for the perception of emotions. Furthermore, Liscombe [12] explored the usefulness of categorical intonation labels in emotion classification in a subset of the EPSAT corpus [11], in which actors read 4-syllable semantically neutral phrases (numbers) in 15 emotions. Ten of the emotions balanced for valence (angry, anxious, bored, confident, encouraging, friendly, frustrated, happy, interested, and sad) plus a neutral utterance from 4 speakers (N=44 utterances) were then subsequently selected for rating of perceived level of emotion by 40 subjects. The relationship between ToBI labels and perceived emotion was then analyzed. Liscombe found a significant effect of pitch accent type on emotion rating for confident, happy, interested, friendly, and bored with L+H* accent showing the greatest disambiguation potential favoring the first five emotions and in general positive valence and high activation while disfavoring boredom. A plateau contour (H*H-L%) was associated with boredom and, in general, H-L% boundaries tended to be associated with negative affect while low boundaries (L-L%) were not.

In addition to the potentially relevant role of phonological description of contour, some studies on very short phrases have shown that knowledge gained from phonological coding of F0 is comparable to information obtained from phonetic features simulating this coding. For example, Benus et al.[4] reported correlation between the backchannel function of cue words and rising F0, which was reflected both in linguistic ToBI labels (H-H% boundaries) and in (stylized) pitch slope extracted over the entire token, its second half, or the last 200ms. These findings suggest that ToBI-like labels offer a feasible description of F0 contours for general analyses, as well as for data with low-quality audio signal.

Our interest in categorical representations of emotion in voice is also motivated by the huge success of such an approach in facial analysis of emotion. For emotion recognition in face, Darwin proposed that a cluster of discrete facial configurations (action units) (such as nostrils raised, mouth compressed, furrowed brow, eyes wide open) need to occur to facially express and interpret emotions [23]. His ideas were later further developed and widely promoted in Ekman's influential work [24], leading to the development of accurate systems for automatic detection of action units on the face [25]. Our interest is in asking if discrete "action units" in voice may similarly be related to emotion expression. A long-term goal would be to develop systems to detect such action units in voice and use them as the basis for emotion prediction, possibly in combination with standard low-level acoustic descriptors.

In this paper we present experiments on a dataset of 433 utterances conveying five basic emotions and manually annotated prosody in the ToBI framework. We analyze the distribution

of discrete prosodic labels in the emotion classes and find several prosodic markers of emotion. Then we present a series of machine learning experiments based on discrete representations of prosody, which have not been done in prior work where the dataset was usually too small to allow for learning. Finally we compare the predictive power of categorical ToBI features and conventional acoustic features, and further combine these two types of features together.

2. Emotional Data

Our data contains recording of emotional utterances produced by 91 professional American actors acting pre-selected sentences under the supervision of a director. The actors were asked to act out a given sentence in a specific target emotion until the director's approval. The dataset contains examples of six basic emotions (*anger, disgust, fear, happy, neutral, sad*) [8] on the following three sentences: [TAI] *The airplane is almost full.*; [TIE] *That is exactly what happen.*; [ITH] *I think I have a doctor's appointment.*

Each of the recorded utterances was classified by ten subjects as expressing one of the possible emotions. The dataset we analyze in this paper contains only utterances which were clearly recognized as the intended emotion by more than five of the of ten raters. We aimed to select 25 examples for each of the six emotions, for each sentence. If more than 25 candidates remained after the perception test, we selected the 25 examples with the highest human recognition rate. Otherwise all validated clear utterances were selected.[1]

3. Annotation of ToBI Labels

All stimuli for a given carrier sentence were randomized across emotions and no information about the emotion label was preserved. They were labeled by an experienced qualified annotator with linguistic and phonetic background, using the ToBI framework [3, 2] within Praat [5].

ToBI labels encode the underlying phonological representation of an utterance primarily in terms of perceived pitch targets (H)igh and (L)ow and disjunctures between words (breaks 0-4 from minimal to strong). Perceptually prominent syllables, primarily due to pitch excursions but also lengthening and intensity, are associated with pitch accents that could consist of single tonal targets (H, !H*, L*), or bi-tonal combinations, most commonly L+H*, L*+H, H+!H*; !H represent a target downstepped from a preceding H target, and "*" corresponds to the tone aligned with the stressed syllable. To provide a rough estimate of pitch range, ToBI includes HiF0 label that indicates the point of the highest F0 in a given intermediate or intonational phrase. For prosodic chunking, breaks 0 and 1 correspond to regular fluent word transitions, 2 to a perceived disjuncture with no salient tonal marking, 3 marks an intermediate phrase associated with H-, L-, or !H- targets, and 4 signals the strongest disjuncture corresponding to the intonational phrase boundary representing a combination of the three phrase accents, marked with dashes and listed above, and a L% or H% tone: H-H%, L-H%, etc. Only minimal adjustments to the ToBI guidelines were employed. Regarding the break indices, diacritics describing uncertainty and disfluency were not used since the nature of the data elicitation minimized these phenomena and we wanted to mitigate data sparsity. Additionally, break 0 was not used. The full set of tonal events was employed.

[1] Specifically there were 23 sad samples of TAI, 19 happy for TIE, 24 disgust and 17 sad for ITH. All other emotions and sentences are represented by 25 different stimuli each.

4. Pitch events across emotions

We now turn to our utterance-level analysis of categorical prosody cues for emotional speech. We consider five sets of ToBI features: types of breaks, intermediate phrase accents, boundary tones in the end of the utterances, pitch accents, and the position of HiF0 in the utterances. For each set of features, we compute the distribution of possible pitch events across different emotion. The results are shown in Table 1. For ease of comparison and interpretation, the first column lists the distributions on *neutral* speech, while the last column gives the average across six emotions. These can be interpreted as benchmarks, providing an insight on how speech associated with a particular emotion differs from neutral speech and the overall average.

The table shows that different emotion classes have different distribution of ToBI features. *Anger* and *disgust* utterances are more likely to contain salient disjunctures between words as shown by higher frequency of 2 and 3 breaks. In particular, *anger* and *disgust* utterances have higher occurrence of !H-L% and L- in the middle of utterance, respectively. Similarly, *fear* and *sad* are characterized by increase in the use of H-L%. *Neutral* utterances on the other had contain only H- intermediate phrasal tones. Hence, emotions with negative valence and high activation tend to be associated with utterances 'chunked' into more prosodic units despite their relatively short duration. This finding is in line with findings that perceived negativity of the word 'whatever' correlated with a prosodic boundary between the first two syllables [4]. Similarly, plateau boundary tones correlate with negative valence, which corroborates the findings discussed in section 1 [12].

Most of the examples in our data are uttered with declarative ending of L-L% final boundary tones. The exceptions are *fear* and *happy* and to a lesser extent also *sad*. They have fewer declarative endings of L-L% and are characterized by a somewhat higher rate of !H-L% and H-L% boundary tones. The association of !H-L% boundary with a positive valence *happy* has not been previously reported and is connected to emotional meaning of H*!H-L% contour that we discuss later.

Compared with *neutral*, emotional speech has more accented words, particularly for *anger*. We can also see the differentiation of emotions on pitch accent distributions. Apart from *neutral*, all other emotions have less occurrence of downstepped (!H*) pitch accents. *Fear* and *sad* have extremely high frequency of H* accents, while *happy, disgusted* and *fear* are associated with the use of L+H* accents. The L+H* association with *happy* and *fear* is a novel surprising finding given the observation of Liscombe's dataset [12] in which mostly positive valence emotion were associated with this accent.

Finally, emotions are very different in terms of the placement of the highest F0 in the prosodic phrase. For example, most of *neutral* and *disgust* utterances start with high pitch, while *fear* and *happy* utterances tend to have highest F0 in the end of the utterances. This HiF0 placement in these two emotion is another indication for the special status of extra-high last pitch accent associated with (L+)H*!H-L% contour for these two emotions, which we discuss later.

In addition to the analysis of individual ToBI labels, we also explore bi-grams of pitch accents. These capture the intonation contour of two concatenated accented words. Final boundary tones were also involved in the bi-gram analysis. Table 1 also lists the most frequent bi-grams in terms of pitch accents and

Table 1: *Distribution of ToBI labels on neutral (first columns) and emotional speech for different types of pitch events and pitch event bigrams*

	NEU	**ANG**	**DIS**	**FEA**	**HAP**	**SAD**	Ave.
breaks (%)							
b1	76.0	68.0	69.8	79.3	78.6	74.1	74.2
b2	2.0	7.2	6.6	0.5	0.5	0.9	3.0
b3	3.4	3.7	3.3	0.7	2.4	3.0	2.8
b4	18.6	21.1	20.3	19.5	18.5	22.0	21.0
phrasal tones – intermediate (%)							
proportion of utts have intermediate phrasal tones							
prop.	20.0	34.7	25.7	8.0	14.5	29.2	21.9
L-L%	0.0	3.8	15.8	0.0	0.0	10.5	6.3
!H-L%	0.0	26.9	5.3	0.0	10.0	0.0	9.5
H-L%	0.0	0.0	10.5	33.3	0.0	26.3	9.5
H-H%	0.0	15.4	0.0	16.7	0.0	10.5	7.4
H-	100.0	38.5	26.3	33.3	60.0	15.8	43.2
!H-	0.0	7.7	0.0	16.7	10.0	26.3	9.5
L-	0.0	7.7	42.1	0.0	20.0	10.5	14.7
boundary tones – end of utts (%)							
L-L%	96.0	97.3	82.4	64.0	68.1	72.3	80.4
!H-L%	0.0	2.7	8.1	21.3	20.3	6.2	9.7
H-L%	1.3	0.0	6.8	10.7	5.8	15.4	6.5
L-H%	2.7	0.0	1.4	2.7	4.3	1.5	2.1
H-H%	0.0	0.0	0.0	1.3	1.4	1.5	0.7
pitch accent (%)							
proportion of accented words							
prop.	50.6	56.8	51.4	53.5	51.5	51.8	52.6
H*	45.1	57.1	47.3	71.5	53.8	66.1	56.8
!H*	35.3	21.6	20.7	12.1	12.3	16.1	19.8
L+H*	9.8	20.4	22.7	11.6	25.7	5.7	16.2
L*	2.5	0.0	4.9	0.0	3.6	4.0	2.4
H+!H*	7.4	0.9	3.9	4.7	4.6	6.9	4.6
occurrence of HiF0 at different position of utts (%)							
begin	73.9	46.5	67.8	38.8	23.7	60.3	52.4
middle	18.2	32.3	21.8	17.5	25.0	16.7	22.2
end	8.0	21.2	10.3	43.8	51.3	23.1	25.4
bigrams - pitch accent (top 10 tokens) (%)							
H*, !H*	28.1	18.5	13.3	13.6	8.3	19.3	16.9
H*, H*	15.6	28.4	17.8	50.7	32.6	31.1	29.4
!H*, !H*	13.3	3.1	4.4	2.1	0.8	2.5	4.4
H*, H+!H*	7.4	0.6	3.7	4.3	1.5	6.7	4.0
L+H*, !H*	7.4	8.6	10.4	2.1	9.1	0.0	6.3
!H*, H*	3.7	9.3	7.4	7.1	3.8	8.4	6.6
L+H*, H*	3.0	9.9	4.4	4.3	6.1	2.5	5.0
H*, L*	1.7	0.0	5.0	0.0	0.0	2.5	1.5
H*, L+H*	0.7	7.4	8.1	5.0	7.6	0.8	4.9
L+H*,L+H*	0.7	4.9	5.2	3.6	6.1	0.0	3.4
bigrams - pitch accent with final boundary tones (top 5) (%)							
H*, L-L%	21.3	50.7	29.7	46.7	30.4	38.5	36.3
!H*,L-L%	56.0	37.3	23.0	6.7	18.8	15.4	26.6
L+H*,L-L%	1.3	5.3	1.4	8.0	13.0	3.1	5.30
H*, !H-L%	0.0	1.3	4.1	14.7	18.8	6.2	7.40
H+!H*,L-L%	12.0	2.7	2.7	2.7	0.0	9.2	4.80

the combination of a pitch accent and a boundary tone.

Compared with the analysis of individual ToBI features, the bi-gram features give us insights of more global and contextual intonation patterns associated with emotions. For example, *fear* can be characterized by increased occurrence of two adjacent H* accents, *happy* and *fear* are associated with pitch accent of H*, followed by down-stepped !H-L% boundary in the end, while *anger* utterances are more associated with H* followed by low (L-L%) boundary. In general, bigrams support the observation about the tendency for avoiding down-stepped pitch accents in emotion speech compared to emotionally neutral speech. One note regarding H*!H-L% contour is that the target for H* pitch accent was commonly extremely high, which gave a particular contour especially for *happy* utterances. Since standard ToBI does not have a dedicated label for this situation, we could not determine if disambiguation between *happy* and *fear* might be facilitated with this additional information. In future work we can consider adaptations of ToBI for emotional speech which would account for this difference.

5. Classification with ToBI features

Our analysis clearly indicates that different emotions exhibit differences on discrete intonation patterns.Now we turn to investigate the effectiveness of these ToBI features for automatic emotion classification. No prior work has tested the applicability of ToBI labels for automatic prediction.

Each utterance is represented by 41 discrete prosodic features which are a combination of all sets of individual ToBI features and bi-gram features in Tables 1. We use SVM classifiers with radial basis kernel constructed using the LIBSVM library [7]. We performed one-versus-all emotion classification tasks: recognition of each of the six emotions versus the other five emotions. For example, one of the tasks was to recognize if an utterance conveys *anger* versus some other emotion among *disgust, fear, happy, sad,* and *neutral*. Since the number of utterances for class *all* is much higher than the one for the target emotion, we performed down-sampling to equal size classes.

To investigate the effect of the sentence structure and context information, we performed experiments on both within-sentence and cross-sentence classification. For the within-sentence classification task, one-versus-all emotion recognition was performed on each of the three selected sentences separately. We perform 10-fold cross-validation on all renditions of the same sentence in different emotions. There are about 145 renditions of each sentence. The accuracy of prediction for each sentence is shown in the top section of Table 2.

Table 2: *One-versus-all accuracy for ToBI features*

	NEU	**ANG**	**DIS**	**FEA**	**HAP**	**SAD**	Ave.
within-sentence classification rate (%)							
TAI	82.4	81.0	81.4	77.6	80.4	73.9	79.5
TIE	78.8	79.6	76.4	76.8	81.6	75.1	78.1
ITH	82.0	76.2	74.3	76.8	74.8	77.9	77.0
Ave.	81.1	78.9	77.4	77.1	78.9	75.6	78.2
cross-sentence classification rate (%)							
TAI	77.0	69.0	69.8	77.0	75.4	57.9	71.0
TIE	69.3	68.0	68.4	64.8	75.9	55.7	67.0
ITH	75.1	67.2	64.1	66.0	51.5	70.8	65.8
Ave.	73.8	68.1	67.4	69.3	67.6	61.5	67.9

The results show that the sentence-specific ToBI features represent a promising line of research for predicting the target emotion. The performance is reasonable on all emotions for all three sentences, with average classification rate of 78.2%. Interestingly at the same time there is a noticeable performance difference for different sentences. For instance, we obtain much higher classification rate of 80.4% and 81.6% on TAI and TIE respectively for *happy*, while the lowest one on ITH of 74.8%.

In our second set of experiments, we turn to the analysis of the cross-sentence performance of emotion recognition. Here we train the model on all data from two sentences and test on the data from the remaining sentence. The results are shown in the bottom part of Table 2, each row representing results when

the given sentence was used as a test set. Compared to the results of within-sentence prediction, cross-sentence accuracy degrades considerably, as can be expected. In contrast to the within-sentence validation, the average classification rate drops from 78.2% to 67.9% and we observe consistent degradation of around 10% on all emotions. This indicates that the ToBI features are highly related to the carrier sentence.

Neutral was the emotion for which classification rate was highest in both types of experiments. It was also the emotion for which there was least degradation in accuracy between the two types of experiments. This finding suggests that the intonation patterns in terms of ToBI features are more robust on *neutral* utterances than on sentences expressing emotions. The worst performance in both experiments is on *sad*, which further corroborates the complex nature of emotions with low-activation.[2]

6. Classification with Acoustic Features

To compare the prediction power of categorical prosody cues and conventional bag-of-frame acoustic features for emotion classification, we conduct similar 1-vs-all emotion classification experiments with a set of state-of-the-art acoustic features.

We use the openSMILE feature extraction library [9] to obtain a comprehensive set of standard acoustic features. The openSMILE library extracts 26 low-level descriptors including intensity, loudness, F0, F0 envelope, probability of voicing, zero-crossing rate, 12 MFCCs, and 8 LSFs. We also use the first order delta coefficients for these features, as well as 19 summary functions for a total of 988 features.

Table 3 lists the corresponding performance of acoustic features in the within-sentence and sentence-independent emotion classification tasks. In the within-sentence emotion classification tasks with conventional acoustic features achieved average accuracy of 78.5%, which is comparable to the 78.2% we obtained with ToBI features in Table 2. However, the prediction power of acoustic features and ToBI features vary among different emotions. For instance, acoustic features show the highest prediction power on *anger*, while ToBI features work better on differentiation of *neutral* and emotional utterances.

Unlike ToBI features, raw acoustic features appear to be less sensitive to the carrier sentence and lead to practically identical accuracy on within- and cross-sentence prediction.

Table 3: *Classification rate of one-versus-all emotion classification using acoustic features.*

	NEU	**ANG**	**DIS**	**FEA**	**HAP**	**SAD**	Ave.
within-sentence classification rate (%)							
TAI	77.0	90.2	72.6	70.2	68.2	79.3	76.3
TIE	84.6	90.0	72.6	74.6	80.0	84.0	81.0
ITH	76.4	90.0	76.2	76.2	71.8	79.4	78.3
Ave.	79.3	90.0	73.8	73.7	73.3	80.9	78.5
cross-sentence classification rate (%)							
TAI	75.4	89.1	71.4	77.4	79.0	77.2	78.3
TIE	87.7	90.6	73.4	67.6	77.6	77.9	79.1
ITH	81.3	88.8	75.1	69.7	78.4	77.4	78.5
Ave.	81.5	89.5	73.3	71.6	78.3	77.5	78.6

[2]In experiments that fall out of the scope of this paper, we also conducted experiments with automatically derived ToBI annotation using AutoToBI [17]. The classification results were poor, and the most predictive ToBI features were practically never recognized correctly, probably because of their rare occurrence in the non-emotional data on which the system is trained.

Finally, we turn to discuss the combination of acoustic and categorical prosodic cues for emotion classification. We apply early stage feature fusion and train SVM classifiers with one feature vector concatenating ToBI features and acoustic features. We consider the sentence-independent task and list the corresponding emotion classification performance of the fusion classifiers in Table 4. Compared with the best single classifiers with standard acoustic features, we can obtain modest improvement by including categorical prosodic ToBI information, where the final average classification accuracy increases from 78.6% to 79.3%. The combined feature representation does achieve noticeable improvement on *neutral, fear,* and *anger*, which is consistent with what we found in cross-sentence results in Table 2 that ToBI features also show higher prediction power on these three emotion classes.

Table 4: *Classification rate of one-versus-all emotion classification by fusion of acoustic and ToBI features*

	NEU	**ANG**	**DIS**	**FEA**	**HAP**	**SAD**	Ave.
sentence-independent classification rate (%)							
TAI	78.2	89.1	71.0	78.2	78.6	75.6	78.5
TIE	88.5	93.0	73.4	69.3	76.0	77.9	79.7
ITH	83.8	88.8	75.1	71.4	81.3	78.2	79.8
Ave.	**83.5**	**90.3**	73.2	**73.0**	78.6	77.2	79.3

7. Conclusions

We have presented a study of the relationship between prototypical emotion expression and discrete, perceptually based characterization on prosody. Our corpus is much larger than any of those used in prior work addressing a similar question. We show promising results both in descriptive and predictive power of ToBI features. Our findings reveal targets for continuous feature extraction that can capture the relevant prosodic phenomena.

In addition, we compare the prediction power of categorical prosodic ToBI features with state-of-the art bag-of-frames acoustic features. We find that ToBI features and acoustic features show comparable performances on within-sentence emotion predictions. Unlike ToBI features, conventional acoustic features are very robust to different carrier sentences. It is nevertheless notable that the much smaller and interpretable set of ToBI features conveys rich information about emotional state.

Finally, we achieved further improvements by integrating ToBI features and acoustic features. This suggests that categorical prosodic ToBI representations can provide complementary information to the conventional acoustic features in prediction of emotion. Our work presents evidence that discrete characterizations of intonation have the potential to inform future feature development for emotion recognition and may lead to overall improved performance.

8. Acknowledgment

This work was supported in part by the following grants: NIH R01-MH073174, NIH P50-MH096891, and NIH R01-MH084856. Some work results from project implementation: Research and development of new information technologies for forecasting and mitigation of crisis situations and safety, ITMS 26240220060 supported by the Research & Development Operational Programme funded by the ERDF.

9. References

[1] Bänziger, T., Scherer, K., "The role of intonation in emotional expressions," *Speech Communication*, vol. 46(3-4), pp. 252-267, 2005.

[2] Beckman, M. E., Ayers, E., Guidelines for ToBI labelling, version 3.0, 1993.

[3] Beckman, M. E., Hirschberg, J., Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework", In S.-A. Jun [ed.], *Prosodic Typology – The Phonology of Intonation and Phrasing*, 2005.

[4] Benus, S., Gravano, A., Hirschberg, J., "The prosody of backchannels in american english," in *Proceedings of ICPhS*, pp. 1065-1068, 2007.

[5] Boersma, P., Weenink, D., Praat: doing phonetics by computer [Computer program], Version 5.3.42, http://www.praat.org, 2013.

[6] Busso, C., Lee, S., Narayanan, S., "Analysis of emotionally salient aspects of fundamental frequency for emotion detection", *IEEE Trans. Audio Speech Language Process.*, Vol. 17(4) pp. 582-596, 2009.

[7] Chang, C.-C., Lin, C.-J., LIBSVM: a library for support vector machines, http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[8] Cowie, R., "Describing the emotional states expressed in speech", In *Proc. of the ISCA Workshop on Speech and Emotion*, pp 11-18, 2000.

[9] Eyben, F., Wöllmer, M., Schuller, B., "openSMILE: The Munich versatile and fast open-source audio feature extractor", in *Proc. of the International Conference on Multimedia*, pp. 1459-1462, 2010.

[10] Laukka, P., Juslin, P. N., "Similar patterns of age-related differences in emotion recognition from speech and music", *Motivation & Emotion*, vol. 31, pp. 182–191, 2007.

[11] Liberman, M., Davis, K., Grossman, M., Martey, N., Bell, J. Emotional prosody speech and transcripts, Linguistic Data Consortium, Philadelphia.

[12] Liscombe, J., *Prosody and Speaker State: Paralinguistics, Pragmatics and Proficiency*, PhD thesis, Columbia University, 2007.

[13] Mozziconacci, S. J. L., *Speech variability and emotion: production and perception*, Ph.D. thesis, Technical University Eindhoven, 1998.

[14] Mozziconacci, S. J. L., "Prosody and Emotions", 2002.

[15] Nakatani, C., Hirschberg, J. and Grosz, B., "Discourse Structure in Spoken Language: Studies on Speech Corpora", in *Proc. of AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[16] Pang, B., Lee, L., "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol. 2(1-2), 1-135, 2007.

[17] Rosenberg, A., "Autobi ? a tool for automatic tobi annotation", in *Proc. of Interspeech*, 2010.

[18] Russ, J. B., Gur, R.C., Bilker, W. B., "Validation of affective and neutral sentence content for prosodic testing", *Behavior Research Methods*, vol. 40(4), pp 935-939, 2008.

[19] Scherer, K. R., Ladd, D. R., Silverman, K., "Vocal cues to speaker affect: testing two models", *Journal of the Acoustic Society of America*, vol. 76(5), pp 1346-1356, 1984.

[20] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., "Tobi: a standard for labeling English prosody", in *Proc. of ICSLP*, pp. 867-870, 2002.

[21] Taylor, P., "The tilt intonation model", in *Proc. of ICSLP*, pp. 1383-1386, 1998.

[22] Wiebe, J., Wilson, T., Cardie, C., "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, Vol. 39(2-3), pp. 165-210, 2005.

[23] Darwin, C., The expression of emotion in man and animals, New York: Oxford University Press. (Original work published 1872)

[24] Ekman, P., "Facial expression of emotion: New findings, new questions", Psychological Science, vol.3, pp. 34–38, 1992.

[25] Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J., "Fully Automatic Facial Action Recognition in Spontaneous Behavior", in *Proc. of FGR 2006*, pp. 223–230, 2006.