INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Analysis of Multi-Lingual Emotion Recognition Using Auditory Attention Features

Ozlem KALINLI

Sony Interactive Entertainment US R&D, San Mateo, CA, USA
ozlem.kalinli@ieee.org

Abstract

In this paper, we build mono-lingual and cross-lingual emotion recognition systems and report performance on English and German databases. The emotion recognition system uses biologically inspired auditory attention features together with a neural network for learning the mapping between features and emotion classes. We first build mono-lingual systems for both Berlin Database of Emotional Speech (EMO-DB) and LDC's Emotional Prosody (Emo-Prosody) and achieve 82.7% and 56.7% accuracy for five class emotion classification (neutral, sad, angry, happy, and boredom) using leave-one-speaker-out cross validation. When tested with cross-lingual systems, the five-class emotion recognition accuracy drops to 55.1% and 41.4% accuracy for EMO-DB and Emo-Prosody, respectively. Finally, we build a bilingual emotion recognition system and report experimental results and their analysis. Bilingual system performs close to the performance of individual mono-lingual systems.

Index Terms: multilingual emotion recognition, auditory attention features, human-computer interaction.

1. Introduction

Emotion recognition is important for many applications including call centers, virtual agents, video games, and human-machine interfaces. For example, a game can dynamically adapt to emotional state of the player; e.g., in a simplistic case, game can become harder or easier based on player's emotional state. The performance of automatic speech recognition (ASR) systems degrade drastically for emotional speech and knowledge of user's emotion can be used to adapt ASR models or to select appropriate pre-trained models dynamically to improve voice recognition performance.

Automatically identifying the emotion of speaker from a given utterance is a challenging task and it has gained more attention in recent years due to wide range of applications. Majority of work in this are has focused on either extracting relevant features from speech that can represent emotions well or classification of emotions based on the extracted features. Traditionally prosodic features such as pitch, duration and energy have been used together with spectral features like mel frequency cepstral features (MFCC), spectral centroid, etc. for emotion recognition [1, 2, 3, 4, 5]. The most commonly used classifiers for emotion recognition are Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), K Nearest Neighbors (KNN), Decision Trees, Support Vector Machines (SVM), and Neural Networks [5, 6, 7, 8, 1, 9].

In this paper, we focus on multilingual emotion recognition inspired by findings of psychological studies on cross-cultural emotion expression and detection. [10, 11, 12] argue that there are "basic" emotions such as happiness, anger, fear, and sadness and they are universally expressed. [13] investigated cross-cultural decoding accuracy between English and Japanese native speakers and found that there was little difference for detecting emotions of non-sense speech spoken in his/her native language or a foreign language. These studies imply that these basic emotions may be expressed and recognized cross culturally or language-independently [14].

In the literature, emotion recognition has been largely explored; however multilingual emotion recognition hasn't been investigated much. [15] compared mono and multi lingual emotion recognition and demonstrated that multilingual emotion recognition using prosodic features can successfully be applied to English, Slovenian, Spanish,and French acted emotional speech recordings. [16] presented a statistical analysis of prosodic features of multilingual emotional speech for Chinese, English, Russian, Korean, and Japanese and confirmed basic emotion states in multilingual speech can be recognized using simple prosodic features. Recently, [17] reported an analysis of features for building a bilingual emotion recognition for anger detection on German and English data. In [18], we recently demonstrated that a probabilistic linear discriminant analysis (PLDA) model trained on German emotional speech data could improve clustering of emotional speech data in English.

Aforementioned studies motivated us to work a step toward multi-lingual emotion recognition in this paper. In our previous work, we proposed to use auditory attention (AA) features for tone and pitch accent classification in English and Mandarin and showed that auditory attention features outperformed the traditional prosodic features for these tasks [19, 20]. Inspired from that, in this study, we propose to use auditory attention cues for emotion recognition for multiple languages. The auditory attention model is biologically inspired and mimics the processing stages in the human auditory system. A neural network is used to learn the mapping between auditory attention features and emotion classes. First, the effectiveness of auditory attention features is demonstrated by building monolingual emotion recognition systems for English and German on five emotions; namely, neutral, angry, sad, happiness, and boredom. Cross-lingual emotion recognition experiments are conducted and demonstrated that there is commonness between the way emotions are expressed and detected in German and English emotional speech. Inspired from multi-style training in automatic speech recognition systems, a multilingual emotion system is built using German and English data together and showed that it can perform as well as monolingual emotion systems do, but with the benefit of having a single model.

The rest of the paper is organized as follows. The emotion recognition system is described in Section 2. The auditory attention model together with feature extraction is explained in

http://dx.doi.org/10.21437/Interspeech.2016-1557

Section 3, which is followed by experimental results in Section 4. The concluding remarks are presented in Section 5.

2. Emotion Recognition Method

It was shown in [21] and the references therein that speech parameters have specific characteristics for certain emotions around syllables; i.e. pitch contour suddenly glides up to a high level within stressed level and then falls to lower level in last syllable for surprise. Here, we used syllables as landmarks for emotion recognition. To detect syllable boundaries automatically, we used the syllable segmentation method in [22], which uses auditory attention features as well.

A window that centers on a syllable is used in order to extract sound around it. Then, auditory attention features are extracted from these sound segments as explained in Section 3. A three layer neural network (3-NN) is used to learn the mapping between features and emotion classes. We use syllable-level auditory attention features as input to the neural network, and the output returns the class posterior probability $p(c_i|f_i)$ for the i^{th} syllable. Here, f_i is the auditory attention feature, c_i is the emotion class and takes values $c_i \epsilon \{1, 2, ..., N_{emo}\}$, where N_{emo} is the number of emotion classes. Then, the emotion tag for a sentence, C^*, is estimated by computing average of posterior scores over all syllables per sentence for each class and taking the maximum as below:

$$C^* = \arg\max_C \frac{1}{N_{syl}} \sum_{i=1}^{N_{syl}} p(c_i|f_i). \quad (1)$$

In Eq.(1), N_{syl} represents the number of syllables per utterance. Next, auditory attention features and how they are extracted is explained.

3. Auditory Attention Model

The block diagram of the auditory attention model is shown in Fig 1. As stated earlier, the model is biologically inspired and hence mimics the processing stages in the human auditory system. First, the auditory spectrum of the input sound is computed based on early stages of the human auditory system. The early auditory system model used here consists of cochlear filtering, inner hair cell, and lateral inhibitory stages mimicking the process from basilar membrane to the cochlear nucleus in the auditory system [20]. The cochlear filtering is implemented using a bank of 128 overlapping constant-Q asymmetric band-pass filters. For analysis, audio frames of 20 milliseconds (ms) with 10 ms shift are used, i.e. each 10 ms audio frame is represented by a 128 dimensional vector.

Next, multi-scale features, which consist of *intensity* (I), *frequency contrast* (F), *temporal contrast* (T), and *orientation* (O_θ) with $\theta = \{45^o, 135^o\}$, are extracted from the auditory spectrum based on the processing stages in the central auditory system [20, 23]. These features are extracted using 2D spectro-temporal receptive filters mimicking the analysis stages in the primary auditory cortex. Each of the receptive filters (RF) simulated for feature extraction is illustrated with gray scaled images in Fig 1 next to its corresponding feature. The excitation phase and inhibition phase are shown with white and black color, respectively. For example, the frequency contrast filter corresponds to receptive fields in the primary auditory cortex with an excitatory phase and simultaneous symmetric inhibitory side bands. Each of these filters is capable of detecting and capturing certain changes in signal characteristics. For example, the orientation features are capable of detecting and capturing when pitch is raising (orientation with 45^o) or falling (orientation with 135^o) [23]. Due to multi-scale structure of the model, auditory attention features capture rich information and can successfully recognize emotion from speaker's voice without requiring explicit prosodic features.

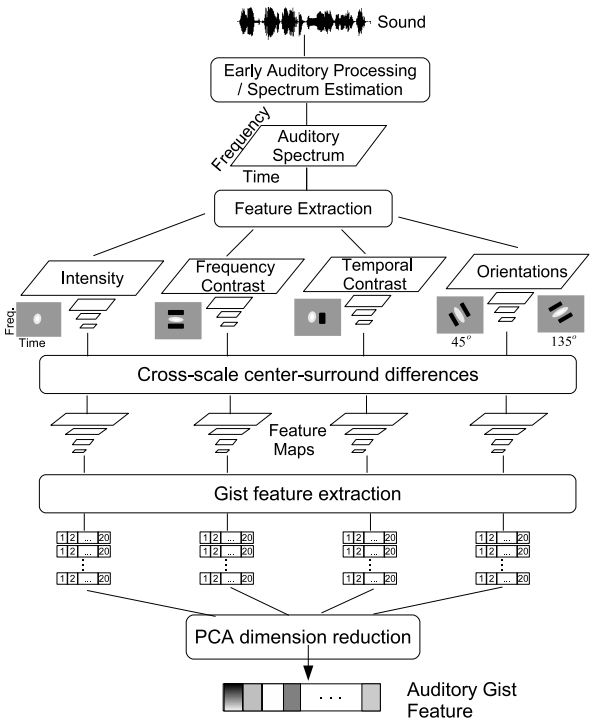

Figure 1: Auditory Attention Model and Gist Extraction

The RF for generating frequency contrast, temporal contrast and orientation features are implemented using 2D Gabor filters with angles 0^o, 90^o, $\{45^o, 135^o\}$, respectively. The RF for intensity feature is implemented using a 2D Gaussian kernel. The multi-scale features are obtained using a dyadic pyramid: the input spectrum is filtered and decimated by a factor of two, and this is repeated. Finally, eight scales are created (if the scene duration is larger than 1.28 seconds (s); otherwise there are fewer scales), yielding size reduction factors ranging from 1:1 (scale 1) to 1:128 (scale 8). For details of the feature extraction and filters, one may refer to [20, 23].

After multi-scale features are obtained, the model computes "center-surround" differences by comparing "center" fine scales with "surround" coarser scales yielding feature maps. The center-surround operation mimics the properties of local cortical inhibition and detects local temporal and spatial discontinuities. It is simulated by across scale subtraction ($\ominus$) between a center scale c and a surround scale s yielding a feature map $\mathcal{M}(c, s)$:

$$\mathcal{M}(c, s) = |\mathcal{M}(c) \ominus \mathcal{M}(s)|, \quad \mathcal{M}\epsilon\{I, F, T, O_\theta\} \quad (2)$$

The across scale subtraction between two scales is computed by interpolation to the finer scale and point-wise subtraction. Here, $c = \{2, 3, 4\}$, $s = c + \delta$ with $\delta\epsilon\{3, 4\}$ are used, which results in 30 feature maps when there are eight scales.

Next, an "auditory gist" vector, also referred here as auditory attention features, is extracted from the feature maps of I, F, T,O_θ such that it covers the whole scene at low resolution.

To do that, each feature map is divided into m-by-n grid of sub-regions and mean of each sub-region is computed to capture the overall properties of the map. For a feature map $\mathcal{M}_i$ with height h and width w, the computation of feature can be written as:

$$G_i^{k,l} = \frac{mn}{wh} \sum_{u=\frac{kw}{n}}^{\frac{(k+1)w}{n}-1} \sum_{v=\frac{lh}{m}}^{\frac{(l+1)h}{m}-1} \mathcal{M}_i(u,v), \qquad (3)$$

for $k = \{0, \cdots, n-1\}, l = \{0, \cdots, m-1\}$. An example of gist feature extraction with $m = 4, n = 5$ is shown in Fig 1, where a $4 \times 5 = 20$ dimensional vector is shown to represent a feature map. After extracting a gist vector from each feature map, we obtain the cumulative gist vector by augmenting them. Then, principal component analysis (PCA) is used to remove redundancy and to reduce the dimension to make machine learning more practical.

4. Experiments and Results

Berlin Database of Emotional Speech (EMO-DB) [24] and LDC's Emotional Prosody Speech Corpus [25] are used in multi-lingual emotion recognition experiments. EMO-DB contains emotional speech from 10 actors (5 female and 5 male) reading 10 German utterances with 7 emotions: anger, boredom, disgust, fear, joy, neutral, and sadness. The Emotional Prosody (Emo-Pro) speech corpus includes English utterances simulated by 7 professional actors[1] by reading short dates and numbers. In Emo-Pro, there are 15 emotion categories: disgust, panic, anxiety, hot anger, cold anger, despair, sadness, elation, happiness, interest, boredom, shame, pride, contempt, and neutral and 2332 utterances in total. In this work, we focus on and present emotion recognition experiments on five common emotion classes: neutral (N), (hot) anger (A), happiness (H), sadness (S), and boredom (B). Since we aim for a multi-lingual and speaker independent system, all experiments are conducted using leave-one-speaker out cross validation. This also provides more fair baseline and comparison for cross-lingual results.

In all experiments, a 3-layer neural network is used for learning the mapping between the auditory attention features and emotion classes. The neural network has D inputs, M hidden nodes and N output nodes, where D is the length of auditory attention feature after PCA dimension reduction when 90% of the variance is retained, and $N = 5$ is the number of emotion classes. Different grid and window sizes are tested for auditory attention feature extraction, and it is found empirically that using a window of $W = 1.0$ s together with 16-by-10 grids is sufficient and performs well in emotion recognition. Then, the size of auditory attention features was computed as $D = 240$ using EMO-DB data. M is varied for each experiment and the best performing one is mentioned in relevant places in following sections. For Emo-Pro, we automatically found syllable boundaries using the method in [22], whereas syllable boundaries that come along with the database are used in Emo-DB experiments.

First, we performed monolingual emotion recognition experiments using EMO-DB and Emo-Pro database and results are presented in Table 1. The number of hidden units are varied from $M = 31$ to 248 as increments of $\times 2$ and $M = 62$ and $M = 31$ was the best performing networks for EMO-DB and Emo-Pro, respectively. Then, the proposed method achieved 56.7% and 82.7% five class emotion recognition accuracy for Emo-Pro and Emo-DB, respectively.

[1]There are eight speakers in Emotional Prosody database but one speaker has only neutral emotion speech; hence it was discarded.

Table 1: Monolingual Emotion Recognition Results

Database	Accuracy
EMO-DB	82.7
EMO-Pro	56.7

Table 2: Cross-Lingual Emotion Recognition Results

Database	Accuracy
EMO-DB	55.1
EMO-Pro	41.4

Table 3: Bilingual Emotion Recognition Results

Database	Accuracy
EMO-DB	78.6
EMO-Pro	54.7

We can compare our results on Emo-Pro with previously published state-of-the-art emotion classification results on the same five class emotion recognition task in [14, 2]. [2] achieved 48.5% accuracy using prosodic and spectral features with a GMM based classifier and improved the performance to 59.5% accuracy by applying speaker normalization. [14] achieved 53% accuracy with pitch, energy, zero crossing features with an HMM for emotion recognition. In summary, emotion recognition results on Emo-Pro presented in this paper compare well and outperform [2] and [14] when no speaker info is available, which demonstrates effectiveness of auditory attention features in emotion recognition task. We cannot compare our results on EMO-DB to prior work, since usually all seven classes in EMO-DB are used in previously reported results.

The confusion matrix for monolingual emotion recognition on English Emo-Pro is presented in Table 4. Note that in all tables, N, A, H, S, and B are used to denote neutral, anger, happiness, sadness, and boredom, respectively. Emotion that has the highest accuracy is anger with 74.8% accuracy. Emotion that has the lowest accuracy is sadness with 38.2% accuracy and it is mostly confused with boredom. The emotion that is confused the most is happiness, and it is with anger (with 30.5% rate), which is a well known paradigm in emotion recognition.

The confusion matrix for monolingual emotion recognition on German Emo-DB is presented in Table 5. Emotion that has the highest accuracy is anger and sadness with 96.8% and 96.7% accuracy, followed by neutral with more than 90% accuracy. Emotion that has the lowest accuracy is happiness with 39.4% accuracy and it is mostly confused with anger. As in English, the emotion that is confused the most is happiness and it is confused with anger (with 54.9% rate).

Second, cross-lingual emotion recognition experiments using mismatched language data are conducted and results are presented in Table 2. Here, the neural network emotion model trained on German Emo-DB is used for English Emotion recognition, and vice versa. With cross-lingual models, 41.4% and 55.1% emotion recognition accuracy is achieved for Emo-Pro and Emo-DB, respectively. As expected, due to mismatched language data, there is 15.3% and 27.6% degradation in Emo-Pro and Emo-DB, respectively, however these numbers are well above chance level[2] for these databases.

[2]Chance level is 25% and 30% for Emo-Pro and Emo-DB, respectively, based on the data distribution.

Table 4: Confusion Matrix for Monolingual Emotion Recognition in Emotional Prosody

	A	H	S	N	B
A	74.8	18.0	4.3	2.2	0.7
H	30.5	45.8	6.8	10.2	0.10
S	11.8	15.1	38.2	9.9	25.0
N	1.3	6.3	16.5	49.4	26.6
B	3.3	11.0	14.3	2.6	68.8

Table 5: Confusion Matrix for Monolingual Emotion Recognition in EMO-DB

	A	H	S	N	B
A	96.8	3.17	0	0	0
H	54.9	39.4	0	5.6	0
S	0	0	96.7	1.6	1.6
N	0	1.3	3.8	91.1	3.8
B	2.5	3.7	8.6	6.2	79.0

Table 6: Cross-Lingual: Confusion Matrix for five emotions in Emotional Prosody using Emo-DB German models

	A	H	S	N	B
A	74.1	23.7	0.7	0.7	0.7
H	46.3	29.9	3.4	11.3	9.0
S	13.2	17.8	32.2	6.6	30.3
N	2.5	12.7	32.9	12.7	39.2
B	9.7	12.3	20.8	8.4	48.7

Table 7: Cross-Lingual: Confusion Matrix for five emotions in EMO-DB using Emo-Pro English models

	A	H	S	N	B
A	64.6	35.4	0	0	0
H	23.3	68.5	4.1	1.4	2.7
S	0	0	81.0	1.7	17.2
N	0	14.3	28.6	13	44.2
B	0	6.5	42.9	1.3	49.4

Also, confusion matrices for cross-lingual models are provided in Table 6 and 7. We see similar trends to monolingual models. For example, anger is the most accurately detected emotion in English and most of the classes are recognized well; i.e. boredom with 48.7%, sadness with 32.2% accuracy. However, it seems like English neutral speech was the most difficult class to recognize when German model is used and it was mostly confused with first boredom, next with sadness in German. Again by looking at the cross-lingual Emo-DB results in Table 7, as with monolingual German model, sadness was the most accurately detected class and least effected one from mismatched language. Most of the emotion classes are detected well in Emo-DB as well; except that German neutral speech was difficult to detect with English emotion models and it was mostly confused with boredom in English. From these experimental results, one may infer that there is commonness between the way emotions are expressed and detected in German and English emotional speech.

Finally, we built bilingual emotion models using both German Emo-DB and English Emo-Pro database and results are presented in Table 3. The number of hidden units are varied from $M = 31$ to 248 as increments of $\times 2$, and $M = 124$ was the best performing network. With bilingual model, 54.7% and 78.6% emotion recognition accuracy is achieved for Emo-Pro and Emo-DB, respectively. It can be concluded that bilingual emotion model performs well and catches the performance of monolingual emotion models without requiring having two separate models. Even though, the model size is approximately doubled with $M = 124$, we can still achieve 50.4% and 74.1% with $M = 62$ if one wants to have a smaller model, which is approximately equal to a monolingual model size.

Table 8: Bilingual Model: confusion Matrix for five emotions in EMO-DB and Emotional Prosody

EMO-PRO

	A	H	S	N	B
A	76.3	19.4	2.2	0.7	1.4
H	32.2	43.5	2.8	10.2	11.3
S	11.8	16.5	38.8	10.5	22.4
N	0	8.9	13.9	45.6	31.7
B	4.6	8.4	14.9	6.5	65.6

EMO-DB

	A	H	S	N	B
A	91.3	8.7	0	0	0
H	45.1	52.1	0	2.8	0
S	0	0	90.3	4.8	4.8
N	0	2.5	2.5	82.3	12.7
B	1.2	1.2	9.9	18.5	69.1

The confusion matrices for Emo-DB and Emo-Pro when bilingual model is used are provided in Table 8. It can be observed that the confusion matrix for each database/language looks similar to the one obtained with monolingual models. For example, anger is the most accurately detected emotion in English with 76.3% accuracy and anger and sadness are the most accurately detected classes in German with more than 90% accuracy. Again, the emotion that is confused the most is happiness and it is confused with anger both in English and German.

5. Conclusion and Future Work

In this paper, we presented a novel emotion recognition approach using auditory attention features and demonstrated its effectiveness via five class emotion recognition experiments with German Emo-DB and English Emotional prosody databases, and compared with previously reported results. The main focus of the paper has been multilingual emotion recognition. We conducted experiments on both German and English emotional databases by building mono-lingual, cross-lingual, and bilingual emotion models. Cross-lingual experiments have shown that there is a commonness between the way emotions are demonstrated and detected in English and German. Also, bilingual emotion model built using both German and English data could perform similar to individual monolingual emotion models.

As part of future work, we would like to include more data from varying languages especially consider tonal languages such as Mandarin to see the effect of tone on multilingual emotion recognition. Also, including facial features can greatly help to achieve a unified multilingual system that performs well across languages and emotions.

6. References

[1] S. M. Yacoub, S. J. Simske, X. Lin, and J. Burns, "Recognition of emotions in interactive voice response systems." in *Proc. of INTERSPEECH*, 2003.

[2] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker dependency of spectral features and speech production cues for automatic emotion classification," in *Proc. of ICASSP*, 2009.

[3] M. Li, A. Metallinou, D. Bone, and S. Narayanan, "Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling," in *Proc. of ICASSP*, 2012, pp. 1937–1940.

[4] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

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

[6] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Proc. of ICASSP*, vol. 2, 2003, pp. II–1.

[7] B. Vlasenko and A. Wendemuth, "Tuning hidden markov model for speech emotion recognition," in *33th Fortschritte Der Akustik*, 2007, pp. 317–320.

[8] 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 *Proc. of ICASSP*, vol. 1, 2004, pp. I–577.

[9] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Computing & Applications*, vol. 9, no. 4, pp. 290–296, 2000.

[10] C. Darwin, P. Ekman, and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[11] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[12] R. Van Bezooijen, *Characteristics and recognizability of vocal expressions of emotion*, 1984, vol. 5.

[13] A. Tickle, "Englishand japanese speakers emotion vocalizations and recognition: a comparison highlighting vowel quality," in *ISCA Workshop on Speech and Emotion*, 2000.

[14] R. Huang and C. Ma, "Toward a speaker-independent real-time affect detection system," in *Proc. of ICPR*, 2006.

[15] V. Hozjan and Z. Kačič, "Context-independent multilingual emotion recognition from speech signals," *International Journal of Speech Technology*, vol. 6, no. 3, pp. 311–320, 2003.

[16] Y. Wang, B. Li, Q. Meng, and P. Li, "Emotional feature analysis and recognition in multilingual speech signal," in *Electronic Measurement & Instruments, 2009. ICEMI'09. 9th International Conference on*, 2009, pp. 4–1046.

[17] T. Polzehl, A. Schmitt, and F. Metze, "Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic/prosodic features for anger detection," 2010.

[18] M. Mehrabani, O. Kalinli, and R. Chen, "Emotion clustering based on probabilistic linear discriminant analysis," in *Proc. of Interspeech*, 2015.

[19] O. Kalinli, "Tone and pitch accent classification using auditory attention cues," in *Proc. of ICASSP*, 2011.

[20] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proc. of Interspeech*, 2007.

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

[22] O. Kalinli, "Syllable segmentation of continuous speech using auditory attention cues." in *INTERSPEECH*, 2011.

[23] O. Kalinli and S. Narayanan, "Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information," *IEEE Trans. Audio, Speech, Lang. Process.*, 2009.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. of Interspeech*, 2005, pp. 1517–1520.

[25] J. Liscombe, J. Venditti, and J. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," in *Proc. of Eurospeech*, 2003.