Interspeech 2018
2-6 September 2018, Hyderabad

Vocalic, Lexical and Prosodic Cues for the INTERSPEECH 2018 Self-Assessed Affect Challenge

Claude Montacié[1], *Marie-José Caraty*[2]

[1]STIH Laboratory, Paris Sorbonne University, 28 rue Serpente, 75006, Paris, France
[2]STIH Laboratory, Paris Descartes University, 45 rue des Saints-Pères, 75006, Paris, France
Claude.Montacie@ParisSorbonne.fr, Marie-Jose.Caraty@ParisDescartes.fr

Abstract

The INTERSPEECH 2018 Self-Assessed Affect Challenge consists in the prediction of the affective state of mind from speech. Experiments were conducted on the Ulm State-of-Mind in Speech database (USoMS) where subjects self-report their affective state. Dimensional representation of emotion (valence) is used for labeling. We have investigated cues related to the perception of the emotional valence according to three main relevant linguistic levels: phonetics, lexical and prosodic. For this purpose we studied: the degree-of-articulation, the voice quality, an affect lexicon and the expressive prosodic contours. For the phonetics level, a set of gender-dependent audio-features was computed on vowel analysis (voice quality and speech articulation measurements). At the lexical level, an affect lexicon was extracted from the automatic transcription of the USoMS database. This lexicon has been assessed for the Challenge task comparatively to a reference polarity lexicon. In order to detect expressive prosody, N-gram models of the prosodic contours were computed from an intonation labeling system. At last, an emotional valence classifier was designed combining ComParE and eGeMAPS feature sets with other phonetic, prosodic and lexical features. Experiments have shown an improvement of 2.4% on the Test set, compared to the baseline performance of the Challenge.

Index Terms: emotional valence, self-reported affect, speech, degree of articulation, voice quality, paralinguistic features, SVM-based detector, Compare challenge

1. Introduction

In the Self-Assessed Affect Challenge [1], speech has to be classified as a valence level (low, medium or high) and should be comparable to that obtained at the self-reported affect [2, 3]. Low valence refers to a negative affect state (e.g., disgust, sadness, fear) while high valence refers to a positive affect state (e.g., elation, joy, happiness). Medium valence refers to a neutral affect state (e.g., calm, tense) or to speech without affect. The Challenge task differs from the usual emotion recognition task for which the affective state of the subject is perceived by another individual. In the Self-Assessed Affect Challenge, aware of their own affective state of mind, the subjects report their state before and after speaking. Significant differences were found between self-assessments of emotion and assessments from outside observers [4, 5, 6, 7]. The awareness of affective state of mind [8] is known to be related to moral judgment [9] and non-rational decision making [10].

Various methods have been developed for the assessment of emotions: electroencephalography [11], self-assessment using verbal scale [12] or non-verbal scale [13], assessment by other using non-verbal behavior such as facial expression [14], posture and gesture [15], verbal behavior such as voice [4] and text-based communication [16].

A vast amount of literature is available about the acoustic and linguistic cues related to emotional speech [17, 18, 19, 20, 21, 22, 23]. The emotional speech databases were frequently based on play-acted speech. Comparative studies between spontaneous emotion, emotion induced, and play-acted emotion [20] do not show strong vocal changes. In this challenge, the goal is to find relevant cues related to the self-assessed affect emotional valence. In related work, the search of acoustic cues related to emotional valence is shown to be difficult and give weak results in discriminating the positive and negative valence [24, 25]. It seems hard to estimate valence from prosodic parameters [24]. In [25], no significant prosodic changes were found between happy (positive valence) and unhappy (negative valence). Various sets of prosodic features have been studied for the emotional valence recognition. Majority of approaches have used the statistics (min, max, mean, standard deviation, skewness, ...) of pitch, energy and speech rates. However some studies [26, 27, 28, 29, 30] have reported that some parameters related to voice quality and speech articulation may be reliable with respect to identification of emotional valence: NAQ (Normalized Amplitude Quotient) [29], harmonic structure of the speech [30] such as HNR (Harmonic to Noise Ratio), Fraction of Locally Unvoiced Frames (FLUF) [27], glottis features [27], F2 and F3 formant values [26, 29]. In a lexical approach [21, 31], lexical cues related to emotional valence seemed more effective to discriminate emotional valence: Continuous Bag-Of-Words (CBOW) [31], Pointwise Mutual Information (PMI) [21], Term Frequency-Inverse Document Frequency (TFIDF), and polarity lexicon. For these experiments, transcription of speech was obtained by manual speech transcripts [21] or by Automatic Speech Recognition (ASR) system [19].

The automatic detection of emotional valence from speech can indicate the level of satisfaction of a person, and provide clues to the social receptivity of a product or decision. It could be a key for novel human-computer interaction applications such as social robotics. The majority of studies [23] on the automatic detection of emotional valence have been conducted over the past decade. Many emotional databases [32, 33, 34] have been developed and provided during challenges [35]. Data-driven approaches were used for all this research. Three kinds of cues were used: two of them were computed from the speech signal (prosodic and spectral cues), and the third one was modeled from the analysis of textual content.

10.21437/Interspeech.2018-1331

For the Interspeech 2018 Self-Assessed Affect Challenge and on the basis of related work, we paid a special attention to the acoustic and linguistic cues that should impact classification performance. In particular, we investigated the voice quality, the speech articulation and the polarity of the words well-known to be sensible cues to emotional valence. The paper is organized as follows: the statistics on the self-Assessed Challenge Corpus (USoMS database) are given in Section 2. Our Basic System (BS) is described in Section 3. In Section 4, features related to voice quality and speech articulation are estimated on vocalic segments and assessed on the Development set. In Section 5, the computation of emotional features related to the lexical content provided by a voice dictation system is described. Prosodic features are studied in Section 6, a stochastic modeling with intonation contours of the audio files is defined and assessed on the Development set. The last section concludes the study.

2. Speech material

The Ulm State-of-Mind in Speech database (USoMS) is used for the Self-Assessed Affect Challenge [1]. The USoMS database consists of a set of 2,113 audio clips (8 second duration) uttered by 100 students (85 females, 15 males). The gender information is missing in the database metadata and the language is German. We developed a gender detection system using cluster analysis of the Train set followed by a listening assessment. A two-class classifier (Male, Female) was trained using eGeMAPS feature set [36] and assessed on the Devel set with an Unweighted Average Recall (UAR) of 99%. For each audio file of the Training (Train) and Development (Devel) sets, the low/medium/high labeling is provided. There are no metadata available on the Test set.

Table 1. *Statistics on the USoMS database.*

Corpora	Train	Devel	Test
# of audio files valence (l, m, h) (female, male)	846 (95, 388, 363) (763, 83)	742 (79, 310, 353) (675, 67)	725 ? (651, 74)
# of phonemes av.: min-max	76.4:3-119	77.0:12-121	78.5:17-116
# of words av.: min-max	16.6: 1-29	16.9: 3-30	16.9: 2-29

Table 1 gives some characteristics and statistics of the USoMS database on the Train, Devel and Test sets. low/medium/high and female/male audio files are significantly unbalanced. Automatic gender detector has been used to discriminate female/male audio files. The statistics (average: minimum-maximum) computed from the number of phonemes and words have been obtained from an automatic speech transcription.

3. Basic system

We have chosen for the development of the emotional valence classifier an SVM-classifier with the combination of ComParE (6,373 features) [36], eGeMAPS (180 features) [36] audio-feature sets and the gender metadata (1 feature). Features were extracted from the audio files using the open source openSmile [37]. Support Vector Machines (SVM) classifier with linear Kernel and Sequential Minimal Optimization (SMO) [38] were used for the emotional valence prediction. To account for the imbalanced class distribution of low valence, the low valence class was up-sampled by a factor of 3 using the SMOTE method. The performance of this classifier in terms of UAR is 57% on the Devel set compared to 56.5% of the official baseline (ComParE functionnals + SVM) [1].

4. Voice quality and speech articulation

Voice quality and speech articulation may be reliable for the identification of emotional valence [26, 28, 29, 30]. Most studies on voice quality have been conducted in the field of speech-language pathology. “Voice quality” refers to the quality of sound produced with a particular shape and tension of the vocal folds such as normal, breathy, creaky, harsh, and tense voices. Articulatory information can be provided directly by methods such as magnetic resonance imaging typically used in Medicine research. Various parameters related to voice quality can be extracted from speech such as sustained vowels: measures of regularity (jitter, shimmer), of breathiness (HNR), of effectiveness (e.g., Soft Phonation Index, NAQ) and parameters of the articulatory model such as the glottal flow relaxation coefficient. Speech articulation refers to Lindblom's theory of Hyper and Hypo-articulation. Hyper-articulated speech is defined as the production of speech with an increase of the articulatory efforts compared to neutral speech (e.g. reading aloud a text emotionless); conversely, Hypo-articulation is defined by the production of speech with minimal articulatory efforts. Hyper-articulated speech is the speaking style usually adopted by a speaker for enhancing speech clarity in difficult communication situation but also to communicate positive emotional valence under certain conditions [39]. One of the effects of hyper-articulation relatively to hypo-articulation is an expansion of the acoustic vowel space constituted by the two first formant frequencies [26]. This expansion makes more distant acoustic targets in the vocalic space explaining partially a higher intelligibility of hyper-articulated speech in comparison with hypo-articulated speech. We chose to compute audio features related to voice quality and speech articulation from vocalic segments. These segments were obtained from the acoustic-phonetic decoding of the speech files.

4.1. Acoustic phonetic decoding

The transcription of the corpus was obtained by an acoustic-phonetic decoding system using a <phone | pause> loop search. The ASR system was based on the version 0.8 of the Pocketsphinx recognizer library [40]. The acoustic models were the pre-trained generic German acoustic models provided by CMU [40].

Table 2. *Percentage of audio clips of the USoMS database containing the phoneme Ph.*

Ph.	ə	ɐ	a	i	ɛ	aɪ	u	o
Occ %	99.7	96	99	98	98	85	88	75

Table 2 gives some statistics of the results of the ASR system on the whole USoMS database. The line gives the percentage of the 2,313 audio clips containing at least a given vowel (e.g., 99.7% of audio clips contain the vowel /ə/).

4.2. Affective vocalic features

From related work reported in the introduction of the section, quality of voice and articulation parameters were chosen for

their impact on emotional valence. For the quality of voice: jitter, shimmer, HNR parameters and FLUF are the four audio features we computed using Praat on all the vocalic segments where they are best estimated. The vocalic space usually measured in German is the vocalic triangle area defined by the three cardinal vowels /a, i, u/ [41]. We chose nine features for the estimation of the articulation parameters related to the vocalic space.

Let us consider Fn_/ph/ the value of the n^{th} formant of the phoneme /ph/ and deltaFn(ph1, ph2) the difference between the n^{th} formant of the phonemes /ph1/ and /ph2/, the nine features are the following: F1_/a/, F2_/a/, F1_/i/, F2_/i/, F1_/u/, F2_/u/, deltaF1(/a/, /i/), deltaF2(/u/, /i/) and the vocalic space area. For the audio clips that do not contain the three cardinal vowels (347/2,133 clips), the vocal space area is labeled undefined.

The relevance of all the thirteen features is given by the information gain IG [42] which is computed on the Train set with the following formula:

$$IG = H(class) - H(class/feature) \quad (1)$$

Where Shannon entropy H is estimated from a table of contingency and class = {low, medium, high}. Features for which IG is greater than zero are considered as relevant.

Nine features out of eleven are relevant and added to our basic system for the emotion valence classification. The ranking order of relevance is the following: F1_/a/, HNR, FLUF, shimmer, deltaF1(/a/, /i/), vocalic space, F1_/i/, jitter, F2_/i/.

These results are consistent with the following three studies [26, 29, 30] that show the sensitivity of vowels to emotional valence on two points: the frequency of the first formant of the phonemes /a/ and /i/ and the harmonic structure of vowels (HNR and FLUF). The performance for this classifier in terms of UAR is 59.5% on the Devel set compared to 57% for our basic system.

5. Affect and polarity lexicon

Lexical cues can be used to effectively discriminate the emotional valence from methods such as the Continuous Bag-Of-Words (CBOW) [31], the Pointwise Mutual Information (PMI) [21], the Term Frequency-Inverse Document Frequency (TFIDF), and the polarity lexicon [43]. Transcripts of audio clips are not available and were obtained through a voice dictation system. This ASR was based on the version 0.8 of the Pocketsphinx recognizer library [40]. The acoustic models are the same as described in Section 4.1. Phonetic lexicon (30,657 words) and language model (trigrams) were provided by CMU [40].

Table 3. *Lexical statistics on the USoMS database*.

Corpora	Train	Devel	Test	All
# of words	14,043	12,540	12,252	38,835
vocabulary size	2,859	2,630	2,596	4,977

Table 3 gives some statistics of the results of the voice dictation system on the whole USoMS database. The first line gives the number of words recognized for each set (Train Devel, Test and the whole database). The second line gives the number of the different recognized words (vocabulary size). After analysis of the lexicons (Train and Devel) we notice that the Devel lexicon differs strongly from the Train lexicon. Many words of the Devel set (1,244/2,630) have no occurrence in the Train set.

5.1. Affective lexical features

We have chosen as affective lexical-features the polarity score of the text transcription of the clip. This score was computed from the polarity of each word of the transcription. Two approaches of word scoring have been assessed: a predefined scored word list and a computation of the score by a supervised algorithm.

For the first approach, SentiWS lexicon (15,649 positive and 15,632 negative words) [43] has been chosen. We notice that only 419 words of the ASR transcript vocabulary (4,977 words) appear in the SentiWS lexicon. The polarity of a clip is computed as the average of the word polarities. If all the words of a clip have no polarity, the polarity of the clip is equal to zero. Linguistic modifiers such as the words “very” or “never” have not been taken into account for this experiment.

For the second approach, the polarity score of each word of the Train set transcription has been computed using PMI method [21]. For this estimation, we use only the clips labeled low and high emotional valence. Consequently, the polarity of the words occurring only in the clip labeled medium emotional valence is zero. The polarity of the clip is computed from the word polarities as described for the first approach.

The relevance of the two affective lexical features is given by the information gain [42] which is computed on the Train set. Only the affective lexical feature computed from SentiWS lexicon is relevant and was added to our basic system for the emotion valence classification.

The performance of this classifier in terms of UAR is 57.8% on the Devel set compared to 57% for our basic system.

6. Modeling of the intonation contours

The prediction of the emotional valence from prosodic parameters is a difficult task. We have developed a new method based on a stochastic modeling of the intonation contours. We hope that features based of this temporal modeling are more discriminating than the usual statistics (min, max, mean, standard deviation, skewness, ...).

The stochastic model is based on the classic N-gram models [44], widely used for speech recognition. Unigram and bigram grammars were trained on the sequence of intonation labels. INTSINT system (INternational Transcription System for INTonation) [45] was chosen for labeling the intonation contours corresponding to each clip.

6.1. MOMEL/INTSINT intonation modeling

The MOMEL/INTSINT method [45] models intonation of an utterance by encoding the pitch curve into a sequence of labels aiming at the stylization of the curve into typical contours of intonation.

MOMEL (MOdelling MELody) algorithm transcripts the smoothed pitch curve by quadratic spline functions into a sequence of target points. Each target point is defined by a couple (P_i, t_i) designing the value of pitch P_i at time t_i. From this sequence, a reduction procedure gives the targets detected as maximum of relevant local variation in the pitch curve and corresponding to major changes in the intonation contour.

Eight target labels are used for the contour stylization of the pitch curve: -three absolute labels from constant pitch

value, T (Top), M (Medium), B (Bottom), and -five contextual labels from variable pitch values depending on the previous target, H (High: local maximum), U (Up-stepped), S (Same as preceding), D (Down-stepped), L (Low: local minimum). These labels are typical contours for the characterization of the intonation.

6.2. N-Gram modeling of intonation contours

The goal of the N-gram modeling is to estimate the probability of a word in a sequence using the previous N-1 words. The probability of a sequence of words is estimated as the product of the word probabilities. This probability can be seen as a measure of the quality of the language model to predict the sequence of words. It is also a measure of dissimilarity between the sequence of words and the corpus that has been used to train the language model.

One bigram model has been estimated for each class of emotional valence. Each language model has been trained on the intonation contours of the Train set corresponding to its class. The three language models (LM_low, LM_medium and LM_high) have been computed using SRILM toolkit [44]. This toolkit was also used to estimate the perplexity of a sequence of words. The perplexity is equal to $P^{-1/n}$ where P is the probability of the sequence and *n* the length of the sequence [44].

We notice that the three bigram models do not use smoothing to estimate unseen sequence of labels. Indeed, all the combinations of two intonation labels occur in their training corpus. In this case, let L1 and L2 be two successive intonation labels, the probability P(L2/L1) is proportional to the number of occurrences of the sequence (L1, L2) in the training set.

Three prosodic features (P_low, P_medium, and P_high) have been defined using the perplexity estimated from the three bigram models (LM_low, LM_medium, and LM_high). The relevance of these prosodic features is given by the information gain [42] which was computed on the Devel set. All these prosodic features were relevant and were added to our basic system for the emotion valence classification.

The performance of this classifier in terms of UAR is 58.1% on the Devel set compared to 57% of our basic system.

7. Experiments on the Test set

To assess our approach, two audio feature sets have been defined as a combination or selection of audio feature sets. The first one D1 is a composite set of 194 features including the eGeMAPS feature set (180 features) [36], the gender metadata (1 feature), the vocalic feature set (9 features: F1_/a/, HNR, FLUF, shimmer, deltaF1(/a/, /i/), vocalic space, F1_/i/, jitter, F2_/i/), the SentiWS lexical feature, the prosodic feature set (P_low, P_medium, P_high). The second one D2 was obtained by feature selection from D1.

On the Test set, three contrastive submissions are described. The first one uses a combination of D1 and ComParE feature sets and gave an UAR of 66.7%. The second one uses a combination of D2 and ComParE feature set and gave an UAR of 67.3%. For the third submission, the confidence score of the basic classifier have been fused with a Bag-of-Audio-Words-based classifier [1] trained on the low level descriptors of ComParE and eGeMAPS. This fusion gave our best result on the Test set with an UAR of 68.4% corresponding to an improvement of 2.4% compared to official baseline performance of the Challenge.

8. Conclusion

In this paper, vocal, lexical and prosodic cues related to the emotional valence of speech have been investigated. A new emotional valence feature set combining audio and lexical features set related to these cues have been defined, estimated and assessed. Two ASR systems have been developed: an acoustic-phonetic decoding system for the estimation of the vocalic features and a voice dictation system for the estimation of the lexical features. A new stochastic modeling of the intonation contours has been defined for the estimation of the prosodic features. The most relevant features are the vocalic features, followed by the prosodic features. The lexical features gave weak results. One explanation is the low coverage of the recognized words by the polarity lexicon.

An emotional valence classifier, combining two SVM classifiers, has been developed. The first one uses a composite set of 6,567 features combining usual feature sets (ComParE and eGeMAPS) and the emotional valence feature set. The second one uses Bag-of-Audio-Words trained on the low level descriptors of ComParE and eGeMAPS. Experiments have shown an improvement of 2.4% on the Test set compared to the official baseline performance of the Challenge (66.0%).

Future work should include an extension of the polarity lexicon using embedded methods [46] to improve the estimation of polarity score of the clip transcriptions. The intra-speaker and inter-speaker variability of the emotional valence classifier should be also studied.

9. References

[1] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats", in Proceedings INTERSPEECH 2018, ISCA, Hyderabad, India, 2018.

[2] J. A. Russell, and A. Mehrabian, "Evidence for a three-factor theory of emotions", Journal of Research in Personality, vol. 11, n° 3, pp. 273–294, 1977.

[3] A.S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients", in Proceedings of the National. Academy of Sciences, vol. 114, n° 38, pp. E7900–E7909, 2017.

[4] S. Biersack and V. Kempe, "Tracing Vocal Expression of Emotion Along the Speech Chain: Do Listeners Perceive What Speakers Feel?", in ISCA Workshop on Plasticity in Speech Perception, pp. 211–214, 2005.

[5] C. Busso and S. S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others", in Proceedings of Interspeech, pp.257–260, 2008.

[6] K. P. Truong, D. A. Van Leeuwen, and F. M. De Jong, "Speech-based recognition of self-reported and observed emotion in a dimensional space", Speech communication, vol. 54, n° 9, pp. 1049–1063, 2012.

[7] N. Smith, "Relations Between Self-Reported and Linguistic Monitoring Assessments of Affective Experience in an Extreme Environment", Wilderness & environmental medicine, vol. 29, pp. 61–65, 2018.

[8] R. L. Mitchell and L. H. Phillips, "The overlapping relationship between emotion perception and theory of mind", Neuropsychologia, vol. 70, pp. 1–10, 2015.

[9] L. Zhang, M. Kong, and Z. Li, “Emotion regulation difficulties and moral judgment in different domains: the mediation of emotional valence and arousal”, Personality and Individual Differences, vol. 109, pp. 56–60, 2017.
[10] Z. R. Steelman and A. A. Soror, “Why do you keep doing that? The biasing effects of mental states on IT continued usage intentions”, Computers in Human Behavior, 73, pp. 209–223, 2017.
[11] A. Al-Nafjan, M. Hosny, Y. Al-Ohali, and A. Al-Wabil, “Review and Classification of Emotion Recognition Based on EEG Brain-Computer Interface System Research: A Systematic Review”, Applied Sciences, vol. 7, n° 12, pp. 1–34, 2017.
[12] J. D. Mayer, and Y. N. Gaschke, “The experience and meta-experience of mood”, Journal of Personality and Social Psychology, vol. 55, pp. 102–111, 1988.
[13] M. M. Bradley and P. J. Lang, (). Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry, vol. 25, n° 1, pp. 49–59, 1994.
[14] P. Ekman and H. Oster, “Facial expressions of emotion”. Annual review of psychology, vol. 30, n° 1, pp. 527–554, 1979.
[15] H. G. Wallbott, “Bodily expression of emotion”, European journal of social psychology, vol. 28, n° 6, pp. 879–896, 1998.
[16] J. T. Hancock, C. Landrigan, and C. Silver, “Expressing emotion in text-based communication”, in Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 929-932, ACM, 2007.
[17] C. E. Williams, , & K. N. Stevens, “Emotions and speech: Some acoustical correlates”, The Journal of the Acoustical Society of America, vol. 52, n° 4B, pp. 1238–1250, 1972.
[18] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first Challenge”, Speech Communication, vol. 53, pp. 1062–1087, 2011.
[19] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011”, Artificial Intelligence Review, vol. 43, n° 2, pp. 155–177, 2015.
[20] K. R. Scherer, “Vocal markers of emotion: Comparing induction and acting elicitation”, Computer Speech & Language, vol. 27, n° 1, pp. 40–58, 2013.
[21] H. Cao, A. Savran, R. Verma, and A. Nenkova. “Acoustic and lexical representations for affect prediction in spontaneous conversations”, Computer speech & language, vol. 29, n° 1, pp.°203–217, 2015.
[22] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, “Analysis of emotional speech - A review”, in Toward Robotic Socially Believable Behaving Systems, vol. 1, pp.°205–238. Springer, Cham, 2016.
[23] D. Torres-Boza, M. C. Oveneke, F. Wang, D. Jiang, W. Verhelst, and H. Sahli, “Hierarchical Sparse Coding Framework for Speech Emotion Recognition”, Speech Communication, vol. 99, pp. 80–89, 2018.
[24] Y. Li, C. T. Ishi, N. Ward, K. Inoue, S. Nakamura, K. Takanashi, and T. Kawahara, “Emotion Recognition by Combining Prosody and Sentiment Analysis for Expressing Reactive Emotion by Humanoid Robot”, in Proceedings of APSIPA Annual Summit and Conference, 4 pages, 2017.
[25] A. S. Cohen, S. L. Hong, and A. Guevara, “Understanding emotional expression using prosodic analysis of natural speech: refining the methodology”, Journal of behavior therapy and experimental psychiatry, vol. 41, n° 2, pp. 150–157, 2010.
[26] M. Goudbeek, J. P. Goldman, and K. R. Scherer, “Emotion dimensions and formant position”, in Proceedings of Interspeech Association, pp. 1575-1578, 2009.
[27] H. P. Espinosa, C. A. R. García, and L. V. Pineda, “Features selection for primitives estimation on emotional speech”, in Proceedings of ICASSP, IEEE, pp. 5138–5141, 2010.
[28] S. Patel, K. R. Scherer, E. Björkner, and J. Sundberg, “Mapping emotions into acoustic space: The role of voice production”. Biological psychology, vol. 87, n° 1, pp. 93–98.
[29] T. Waaramaa, A. M. Laukkanen, M. Airas, and P. Alku, “Perception of emotional valences and activity levels from vowel segments of continuous speech”, Journal of voice, vol. 24, n° 1, pp. 30–38, 2010.
[30] M. C. Sezgin, B. Gunsel, and G. K. Kurt, “Perceptual audio features for emotion detection”, EURASIP Journal on Audio, Speech, and Music Processing, vol. 1, n°16, 2012.
[31] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost, “Pooling acoustic and lexical features for the prediction of valence”, in Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 68–72, 2017.
[32] T. Bänziger, H. Pirker, and K. Scherer, “GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions”, in Proceedings of LREC, Vol. 6, pp. 15–019, 2006.
[33] M. Grimm, K. Kroschel, and S. Narayanan, “The Vera am Mittag German audio-visual emotional speech database”, in IEEE International Conference on Multimedia and Expo, pp.°865–868, 2008.
[34] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, “The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent”, IEEE Transactions on Affective Computing, vol. 3, n° 1, pp. 5–17, 2012.
[35] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, “AVEC 2012: the continuous audio/visual emotion challenge”, in Proceedings of the 14th ACM international conference on Multimodal Interaction, pp. 449–456, 2012.
[36] F. Eyben, “Standard Baseline Feature Sets. In Real-time Speech and Music Classification by Large Audio Feature Space Extraction”, Springer International Publishing, pp. 123–137, 2016.
[37] F. Eyben, F. Weninger, F. Groß, and B. Schuller, “Recent developments in openSMILE, the Munich open-source multimedia feature extractor,” in Proceedings of ACM MM, Barcelona, Spain, pp. 835–838, 2013.
[38] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: An update,” ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10–18, 2009.
[39] S. Jeannin, C. Gilbert, M. Amy, and G. Leboucher, “Pet-directed speech draws adult dogs’ attention more efficiently than Adult-directed speech”, Scientific reports, vol. 7, n° 17, pp. 49–80, 2017.
[40] A. Chan, E. Gouva, R. Singh, M. Ravishankar, R. Rosenfeld, Y. Sun, D. Huggins-Daines, and M. Seltzer, “The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Relate Resources” www.cs.cmu.edu/~archan/share/sphinxDoc.pdf, 2007.
[41] W. Heeringa, H. Schoormann, and J. Peters, “Cross-linguistic vowel variation in Saterland: Saterland Frisian, low German, and high German”, in Proceedings of the 18th International Conference of Phonetic Sciences, pp. 1041–1045, 2015.
[42] T.W. Rauber, A.S. Steiger-Garcao, “Feature selection of categorical attributes based on contingency table analysis. Portuguese Conference on Pattern Recognition, Porto, Portugal, 1993.
[43] R. Remus, U. Quasthoff, and G. Heyer, “SentiWS-A Publicly Available German-language Resource for Sentiment Analysis”, in LREC, pp. 1168–1171, 2010.
[44] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, “SRILM at sixteen: Update and outlook”, in Proceedings of IEEE Automatic Speech Recognition and Understanding workshop, 5 pages, 2011.
[45] D. Hirst, “Form and function in the representation of speech prosody”, Speech Communication, vol. 46, n° 3–4, pp. 334–347, 2005.
[46] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Sted, “Lexicon-based methods for sentiment analysis”, Computational linguistics, vol. 37, n° 2 , pp. 267–307, 2011.