Salient Cross-lingual Acoustic and Prosodic Features for English and German Emotion Recognition

Maxim Sidorov, Christina Brester, Stefan Ultes, and Alexander Schmitt

Ulm University, Ulm, Germany,
{maxim.sidorov,stefan.ultes,alexander.schmitt}@uni-ulm.de
Siberian State Aerospace University, Krasnoyarsk, Russia
christina.brester@sibsau.ru

Abstract. While approaches on automatic recognition of human emotion from speech have already achieved reasonable results, still there remains a lot of room for improvement. In our research, we select the most essential features by applying a self-adaptive multi-objective genetic algorithm. The proposed approach is evaluated using data from different languages (English and German) with two different feature sets consisting of 37 and 384 dimensions, respectively. The obtained results of the developed technique have increased the emotion recognition performance by up to 49.8% relative improvement in accuracy. Furthermore, in order to identify salient features across speech data from different languages, we analysed the selection count of the features to generate a feature ranking. Based on this, a feature set for speech-based emotion recognition based on the most salient features has been created. By applying this feature set, we achieve a relative improvement of up to 37.3% without the need of time-consuming feature selection using a genetic algorithm.

Keywords: emotion recognition, speech analysis, dimensionality reduction, salient features

1 Introduction

Automatic recognition of human emotions based on the speech signal is in the focus of research groups all over the world. While enabling machines to recognize human emotions may be useful in various applications, e.g., improvement of Spoken Dialogue Systems (SDSs) or monitoring agents in call-centers, the recognition performance is still not satisfying.

Emotion recognition (ER) rendered as a classification problem may be solved with supervised learning approaches by extracting a huge amount of numerical features out of the speech signal. However, the curse of dimensionality [1], i.e., having more features results in worse performance, poses a critical issue. Some of the features may be highly-correlated or their level of variability may be dramatically low. Therefore, some attributes might not bring a beneficial impact to the

system—or even decrease its performance. Hence, feature selection techniques are applied which not simply results in a trade-off between time-consuming feature extraction and the accuracy of the model. Moreover, selecting an optimal feature set potentially increases the overall performance of the system. Furthermore, as feature selection is a time-consuming procedure, it is highly desirable to have a set of salient features which is known to result in good emotion recognition performance for different settings, e.g., for different languages.

However, a suitable feature set should be both, representative and compact, i.e., result in good performance while being as small as possible. Hence, in this contribution, we propose the usage of a multi-objective genetic algorithm (MOGA), which is a heuristic algorithm of pseudo-boolean optimization, in order to maximize ER accuracy and minimize the size of the feature set simultaneously. Furthermore, we proposed a self-adaptive scheme of MOGA which exempts from the necessity of choosing the parameters of the algorithm. Eventually, our main focus lies on identifying salient cross-lingual features for emotion recognition. Hence, we analysed the distributions of selected features of three different databases (English and German) in order to identify the most salient features across different databases and languages. Finally, the resulting feature set is evaluated for each of the databases used for creating the feature set.

The rest of the paper is organized as follows: Significant related work is presented in Section 2. Section 3 describes the applied corpora and renders their differences. Our approach on automated emotion recognition using MOGA-based feature selection is presented in Section 4 having its results of numerical evaluations in Section 5. In Section 6, we analyze the cross-lingual feature set and present the list of salient features. Conclusion and future work are described in Section 7.

2 Significant Related Work

One of the pilot experiments which deals with speech-based emotion recognition has been presented by Kwon et al. [2]. The authors compared emotion recognition performance of various classifiers: support vector machine, linear discriminant analysis, quadratic discriminant analysis and hidden Markov model on SUSAS [3] and AIBO [4] databases of emotional speech. The following set of speech signal features have been used in the study: pitch, log energy, formant, mel-band energies, and mel frequency cepstral coefficients (MFCCs). The authors have managed to achieve the highest value of accuracy (70.1% and 42.3% on the databases, correspondingly) using Gaussian support vector machine.

The authors in [5] highlighted the importance of feature selection for the ER manifested by determining an efficient feature subset was using the fast correlation-based filter feature selection method. A fuzzy ARTMAP neural network [6] was used as an algorithm for emotion modelling. The authors have achieved an accuracy of over 87.52% for emotion recognition on the FARSDAT speech corpus [7].

Table 1. Databases description

Database	Language	Length (min.)	# of emotions	File level duration		Emotion level duration	
				Mean(sec.)	Std. (sec.)	Mean (sec.)	Std. (sec.)
Berlin	German	24.7	7	2.7	1.02	212.4	64.8
SAVEE	English	30.7	7	3.8	1.07	263.2	76.3
VAM	German	47.8	4	3.02	2.1	717.1	726.3

While our research of identifying cross-lingual salient features includes four different emotions, Polzehl et al. [8] only focused on the emotion anger. The authors analysed two different anger corpora of German and American English to determine the optimal feature set for anger recognition. The German database contains 21 hours of recordings from a German Interactive Voice Response (IVR) portal offering assistance troubleshooting. For each utterance, three annotators assigned one of the following labels: not angry, not sure, slightly angry, clear anger, clear rage, and garbage. Garbage marked utterances are non applicable, e.g., contain silence or critical noise. The English corpus originates from an US-American IVR portal capable of fixing Internet-related problems. Three labelers divided the corpus into angry, annoyed, and non-angry turns. A total of 1,450 acoustic features and their statistical description (e.g, means, moments of first to fourth order, the standard deviation) have been extracted from the speech signal. The features are divided into seven general groups: pitch, loudness, MFCC, spectrals, formants, intensity, and other (e.g., harmonics-to-noise). Analyzing each feature group separately, the authors achieve in a baseline approach without further feature selection a maximal $f1$ score of 68.6 with 612 MFCC-based features of the German corpus and a maximal $f1$ score of 73.5 with 171 intensity-based features of the English corpus.

3 Corpora

For identifying salient features for emotion recognition, the following three different speech databases have been used:

Berlin DB The Berlin emotional database [9] was recorded at the Technical University of Berlin and consists of labeled emotional German utterances which were spoken by 10 actors (5 f). Each utterance has been assigned one of the following emotion labels: neutral, anger, fear, joy, sadness, boredom or disgust.

SAVEE The SAVEE (Surrey Audio-Visual Expressed Emotion) corpus [10] was initially recorded for research on audio-visual emotion classification containing four native English male speakers. One emotion label for each utterance has been applied using the standard set of emotions (anger, disgust, fear, happiness, sadness, surprise and neutral).

VAM The VAM database [11] was created at the University of Karlsruhe and consists of utterances extracted from the popular German talk-show "Vera

am Mittag" (Vera in the afternoon). The emotion annotation of the first part of the corpus (speakers 1-19) were given by 17 human evaluators and the rest of the utterances (speakers 20-47) were annotated by 6 raters, all on the 3-dimensional emotional basis (valence, activation and dominance). For this work, only pleasantness (or evaluation) and the arousal axis are used. The quadrants (counterclockwise, starting in positive quadrant, assuming arousal as abscissa) are then assigned to emotional labels happy-exciting, angry-anxious, sad-bored, and relaxed-serene (cf. [12]).

A statistical description of the used corpora is depicted in Table 1. Both, the Berlin and the SAVEE database consist of acted emotions while VAM comprises real emotions. Furthermore, the VAM database is highly unbalanced (see Emotion level duration columns in Table 1).

4 Feature Selection with MOGA

For our main contribution of applying feature selection using an adaptive multi-objective genetic algorithm (MOGA) to identify cross-language salient features, a probabilistic neural network (PNN) [13] has been chosen arbitrarily as a classification algorithm for building emotion recognition models as it has shown to be a fast classification algorithm providing good results.

A MOGA, being a genetic algorithm (GA) implementing an effective pseudo-boolean optimization procedure, is used to solve the multi-objective optimization problem. A multi-objective GA operates with a set of binary vectors coding the subsets of informative features, where *false* corresponds to non-essential attributes and *true* corresponds to essential ones.

In this work, the applied MOGA is based on the Strength Pareto Evolutionary Algorithm (SPEA) [14], where non-dominated points are stored in the limited capacity archive named *outer set*. The content of this set is updated throughout the algorithm execution and as a result we have an approximation of the Pareto set.

The scheme of the SPEA method includes several steps:

1. Determination of the initial population $P_t(t = 0)$.
2. Copying of the individuals not dominated by P_t into the intermediate outer set $\left(\bar{P}'\right)$.
3. Deletion of the individuals dominated by $\bar{P}'$ from the intermediate outer set.
4. Clustering of the outer set $\bar{P}_{t+1}$ (if the capacity of the set $\bar{P}'$ is more than the fixed limit).
5. Compilation of the outer set $\bar{P}_{t+1}$ with the set $\bar{P}'$ individuals.
6. Application of all of the genetic operators: selection, crossover, mutation.
7. Test of the stop-criterion: If it is true, then the GA is completed. Otherwise, continue from the second step.

It is well-known that the performance of conventional GAs completely depends on the settings of the genetic operators (selection, crossover, mutation).

Consequently, to achieve a good performance level with SPEA, its parameter settings have to bee adjusted carefully (step 6). Therefore, SPEA-modification based on the idea of self-adaptation has been developed [15].

In this self-adaption version of SPEA, tournament selection is applied: individuals can be selected both from the current population and from the outer set. Hence, only recombination and mutation operators require adjustment (tuning or control).

The mutation probability p_m can be determined according to one of the rules developed by Daridi et al. [16]. As parametrized by Daridi et al., the following rule was used for the proposed self-adaptive SPEA-modification:

$$p_m = \frac{1}{240} + \frac{0.11375}{2^t} \,, \tag{1}$$

where t is the current generation number.

The self-configurable recombination operator is based on the *co-evolution* idea [17]: the population is divided into groups and each group is generated with a particular type of recombination (it may be *one-point*, *two-point* or *uniform* crossover). The size of the subpopulation depends on the fitness value of the corresponding recombination type, where the fitness value q_i of the i-th recombination operator is determined by

$$q_i = \sum_{l=0}^{T-1} \frac{T-l}{l+1} \cdot b_i, \tag{2}$$

where T is the adaptation interval, $l = 0$ corresponds to the latest generation in the adaptation interval, $l = 1$ corresponds to the previous generation, etc. b_i, indicating the effectiveness of the individuals, is defined as

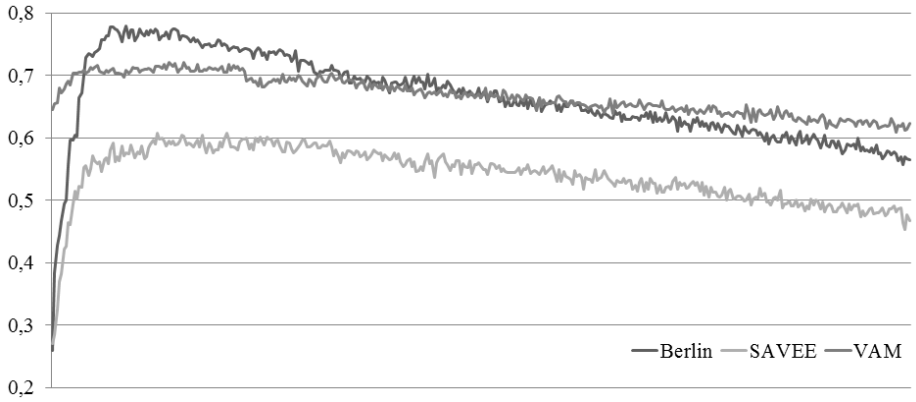
$$b_i = \frac{p_i}{|\bar{P}|} \cdot \frac{N}{n_i} \,, \tag{3}$$

where p_i is the amount of individuals in the current outer set generated with the i-th type of recombination operator, $|\bar{P}|$ is the outer set size, n_i is the amount of individuals in the current population generated with the i-th type of crossover, and N is the population size.

The efficiency of the operators is compared in pairs in every T-th generation to reallocate resources on the basis of the fitness values. Hence, the maximum number of applications $f(n)$ of each individual is defined by

$$f(n) = \begin{cases} 0, & \text{if } n_i \le social\_card \\ int\left(\frac{n_i - social\_card}{n_i}\right), & \text{if } (n_i - h_i \cdot penalty) \\ & \quad \le social\_card \\ penalty, & \text{otherwise} \end{cases} \tag{4}$$

Fig. 1. Accuracy of Emotion Recognition System A: Usage of individual distribution.



Here, s_i is the size of a resource given by the i-th operator to those which won, h_i is the number of losses of the i-th operator in paired comparisons, the $social\_card$ is the minimum allowable size of the population, the *penalty* is a negative score for defeated operators.

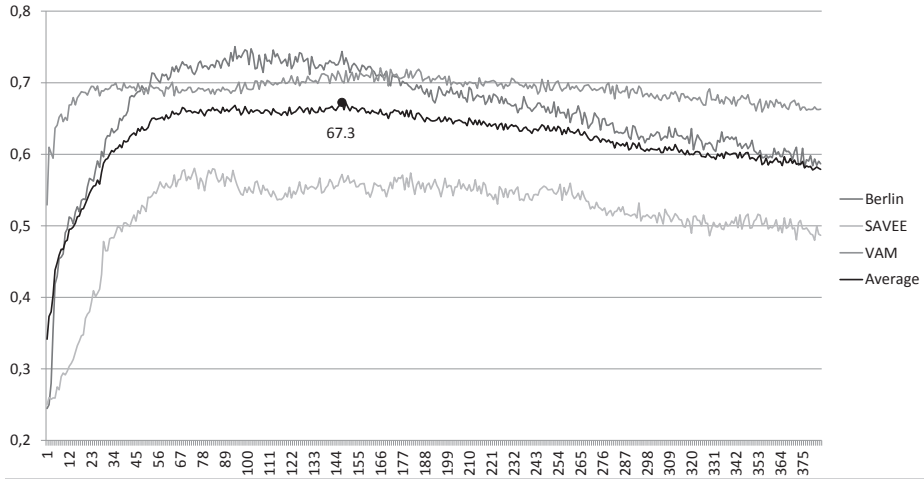
The effectiveness of the proposed approach was evaluated on the set of test problems [18] developed by the scientific community for the comparison of evolutionary algorithms: the effectiveness of the adaptive SPEA outperformed the effectiveness of the average conventional SPEA proving that self-adaptation is an alternative to the random choice of genetic operators or multiple runs of the GA for each variant of settings.

5 Evaluation and Results

To investigate the performance of the MOGA-based feature selection for emotion recognition, two experiments using two different feature sets have been conducted. The first feature set consists of the 37 most common acoustic features and is used as a baseline. Average values of the following speech signal features are included into the baseline feature set: power, mean, root mean square, jitter, shimmer, 12 MFCCs and 5 formants. Mean, minimum, maximum, range and deviation of the following features have also been used: pitch, intensity and harmonicity. All of the 37 features have been extracted for each speech signal file.

The extended feature set consists of 384 features and taken from the Interspeech 2009 Emotion Challenge [19]. It is an extension to the baseline feature set containing additional features and so called functionals. In contrast to the baseline feature set, it is not applied completely but also using the previously presented MOGA feature selection procedure.

Fig. 2. Accuracy of Emotion Recognition System B: Usage of average distribution.



Features of the baseline feature set have been extracted using Praat [20]. The features of the extended feature set have been extracted for the audio signal using the openSMILE toolkit [21].

For evaluating the emotion recognition performance, the data is divided into training and testing set with a ratio of 0.7 for training and 0.3 for testing. The training set is used for creating and training the PNN-based emotion model while the testing set is used for measuring the model's performance. In order to get profound results, this procedure has been repeated fifteen times. The achieved accuracy of emotion recognition with the baseline and the extended feature set is shown in Table 2 (Columns Baseline and IS'09).

For applying feature selection and creating the corresponding emotion model, the training set is used as well. For the feature selection process, the feature set was coded with boolean vectors (*true* corresponds to an essential attribute, *false* to an unessential one) each representing one individual of the self-adaptive MOGA. To test the fitness of the individuals, the training set was in turn divided into two sets of which 80% were used to train a PNN-based emotion model and 20% to evaluate it. As fitness, the average accuracy computed out of the accuracies of fifteen evaluation cycles for each individual was used. After 100 iterations of the MOGA, the optimal feature set, i.e., the fittest individual, was used for creating an emotion model whose performance was then evaluated using the testing set. This complete procedure has also been applied fifeteen times. The results depicted in Table 2 show significant improvement of applying feature selection over the baseline and the extended feature set. As described before, feature selection was only applied for the extended feature set.

Table 2. Evaluation Result of Emotion Recognition: Accuracy with the 37-dimensional feature set (Baseline), the extended feature set (IS'09), and the reduced feature set (GA) having the number of features in parentheses (Num.) and relative improvement of the feature selection approach (Gain)

Database	Baseline	IS'09	GA (Num.)	Gain
Berlin	56.7	58.9	71.5 (68.4)	26.1%
VAM	68.0	67.1	70.6 (64.8)	3.9%
SAVEE	41.6	47.3	48.4 (84.1)	16.3%

Table 3. Evaluation results of the individual and combined ranking (Number of features in parentheses) of the extended feature set. Furthermore, results for the common feature set plus relative improvement (Gain)

Corpus	Individual Ranking	Common Ranking	Common Set	Gain
Berlin	79.7 (86)	75.1 (94)	74.37	31.2%
VAM	73.6 (81)	72.1 (166)	70.24	3.3%
SAVEE	62.3 (148)	58.0 (74)	57.17	37.3%

6 Salient Features for Cross-lingual Emotion Recognition

While feature selection poses a good way of improving the performance, GA-based feature selection is resource-consuming and thus hard to deploy in real-world applications. Hence, in this contribution, we derive the most appropriate feature set based on information about the selected features. The resulting feature set may then be applied to unseen data (language and domain) without the need of running through the complete process anew.

For the creation of optimal feature set for each database, a ranking of the selected feature set has been created for each database separately (Experiment 1). Here, the rank of each feature is based on the number of its participation in the individuals of the GA (feature counts). Then, in order to obtain an optimal feature set for ER in general, a combined ranking of the selected features has been created by combining the feature counts of each database.

For all databases, the information about extracted features was accumulated through multiple executions of the algorithm. Based on these, a distribution of feature selection frequencies (FSF), i.e. the relative number of cases in which a particular feature was selected, was created. Then, the list of features was ranked in the descending order of their FSF-values. During this iterative procedure, the optimal number of features was defined for each database separately and a learning curve was generated by subsequently adding the top-ranked feature to the applied feature set. For providing statistic significant results, the classification procedure was executed 25 times for each set of feature sets. At Figure 1, the accuracy of emotion recognition for all databases is illustrated depending on the number of involved features. It may be noticed that every obtained curve has

Table 4. Top 20 ranked features using self-adaptive MOGA feature selection.

Rank	Feature Group	Rank	Feature Group
1	MFCC	11	MFCC
2	Energy	12	Energy
3	ZCR	13	Energy
4	MFCC	14	Energy
5	MFCC	15	MFCC
6	MFCC	16	MFCC
7	MFCC	17	MFCC
8	MFCC	18	MFCC
9	MFCC	19	MFCC
10	MFCC	20	MFCC

only one maximum and, moreover, the maximums are located in a similar range for all databases.

Based on individual distribution of FSF-values for each database, an combined distribution was derived by calculating the average frequencies. The next steps are similar to Experiment 1: based on the new ranking, the iterative procedure was launched for each database. Figure 2 presents the learning curve, i.e., accuracy of emotion recognition for each iteration. Furthermore, the figure also contains an average learning curve also indicating the combined maximum accuracy with 147 features constituting the common feature set. Again, all values have been determined by taking the average of 25 interactions of the experiment.

Table 3 shows the results for individual and common ranking as well as by applying the common feature set. Individual ranking clearly achieves the best results increasing the performance up to 49.8 % relative to the baseline. However, applying common ranking still provides results clearly above the baseline. Moreover, both results outperform the results of simply applying MOGA-based feature selection as in Table 2. Finally, applying the derived common feature set consisting of the 147 most salient features also results in improvement of accuracy by up to 37.3% relative to the baseline. In Table 4, the top twenty rankend features of the common feature set are shown along with their feature group. When comparing these to the results of Polzehl et al. [8], they belong to similar groups confirming their results for general emotion recongition.

7 Conclusion and Future Work

In this work, we presented the application of a PNN-MOGA hybrid emotion recognition approach where MOGA is used for feature selection. By applying the algorithm to three different corpora containing English or German speech, the overall accuracy could be improved for all data by up to 49.8%. Finally, we created a common feature set out of the individual rankings consisting of the 147 most salient features improving the performance on all corpora by up to 37.3%.

While a PNN has already provided reasonable results for emotion recognition, we still examine its general appropriateness. The usage of other possibly more

accurate classifiers may improve the performance of this system. Furthermore, dialogues may not only consist of speech, but also of a visual representation. Hence, an analysis of video recordings may also improve the ER performance.

References

1. Bellman, R.: Dynamic programming. Princeton University Press (1957)
2. Kwon, O.W., Chan, K., Hao, J., Lee, T.W.: Emotion recognition by speech signals. In: INTERSPEECH (2003)
3. Hansen, J.H., Bou-Ghazale, S.E., Sarikaya, R., Pellom, B.: Getting started with susas: a speech under simulated and actual stress database. In: EUROSPEECH. vol. 97, pp. 1743–46 (1997)
4. Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M.J., Wong, M.: " you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In: LREC (2004)
5. Gharavian, D., Sheikhan, M., Nazerieh, A., Garoucy, S.: Speech emotion recognition using fcbf feature selection method and ga-optimized fuzzy artmap neural network. Neural Computing and Applications 21(8), 2115–2126 (2012)
6. Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., Rosen, D.B.: Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. Neural Networks, IEEE Transactions on 3(5), 698–713 (1992)
7. Bijankhan, M., Sheikhzadegan, J., Roohani, M., Samareh, Y., Lucas, C., Tebyani, M.: Farsdat-the speech database of farsi spoken language. In: the Proceedings of the Australian Conference on Speech Science and Technology. vol. 2, pp. 826–830 (1994)
8. Polzehl, T., Schmitt, A., Metze, F.: Salient features for anger recognition in german and english ivr portals. In: Minker, W., Lee, G.G., Nakamura, S., Mariani, J. (eds.) Spoken Dialogue Systems Technology and Design, pp. 83–105. Springer New York (2011), 10.1007/978-1-4419-7934-6_4
9. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of german emotional speech. In: Interspeech. pp. 1517–1520 (2005)
10. Haq, S., Jackson, P.: Machine Audition: Principles, Algorithms and Systems, chap. Multimodal Emotion Recognition, pp. 398–423. IGI Global, Hershey PA (Aug 2010)
11. Grimm, M., Kroschel, K., Narayanan, S.: The vera am mittag german audio-visual emotional speech database. In: Multimedia and Expo, 2008 IEEE International Conference on. pp. 865–868. IEEE (2008)
12. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.: Acoustic emotion recognition: A benchmark comparison of performances. In: Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on. pp. 552–557. IEEE (2009)
13. Specht, D.F.: Probabilistic neural networks. Neural networks 3(1), 109–118 (1990)
14. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. Evolutionary Computation, IEEE Transactions on 3(4), 257–271 (1999)
15. Eiben, A.E., Hinterding, R., Michalewicz, Z.: Parameter control in evolutionary algorithms. Evolutionary Computation, IEEE Transactions on 3(2), 124–141 (1999)

16. Daridi, F., Kharma, N., Salik, J.: Parameterless genetic algorithms: review and innovation. IEEE Canadian Review (47), 19–23 (2004)
17. Potter, M.A., De Jong, K.A.: A cooperative coevolutionary approach to function optimization. In: Parallel Problem Solving from Nature–PPSN III, pp. 249–257. Springer (1994)
18. Zhang, Q., Zhou, A., Zhao, S., Suganthan, P.N., Liu, W., Tiwari, S.: Multiobjective optimization test instances for the cec 2009 special session and competition. University of Essex, Colchester, UK and Nanyang Technological University, Singapore, Special Session on Performance Assessment of Multi-Objective Optimization Algorithms, Technical Report (2008)
19. Kockmann, M., Burget, L., Černockỳ, J.: Brno university of technology system for interspeech 2009 emotion challenge. In: Tenth Annual Conference of the International Speech Communication Association (2009)
20. Boersma, P.: Praat, a system for doing phonetics by computer. Glot international 5(9/10), 341–345 (2002)
21. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the international conference on Multimedia. pp. 1459–1462. ACM (2010)