Automatic Assessment of Singer Traits in Popular Music: Gender, Age, Height and Race

Conference Paper · January 2011

CITATIONS
26

READS
252

3 authors, including:

Felix Weninger
Nuance Communications
101 PUBLICATIONS **4,989** CITATIONS
SEE PROFILE

Björn Schuller
Imperial College London
1,178 PUBLICATIONS **35,281** CITATIONS
SEE PROFILE

AUTOMATIC ASSESSMENT OF SINGER TRAITS IN POPULAR MUSIC: GENDER, AGE, HEIGHT AND RACE

Felix Weninger, Martin Wöllmer, Björn Schuller
Institute for Human-Machine Communication, Technische Universität München, Germany
(weninger|woellmer|schuller)@tum.de

ABSTRACT

We investigate fully automatic recognition of singer traits, i. e., gender, age, height and 'race' of the main performing artist(s) in recorded popular music. Monaural source separation techniques are combined to simultaneously enhance harmonic parts and extract the leading voice. For evaluation the UltraStar database of 581 pop music songs with 516 distinct singers is chosen. Extensive test runs with Long Short-Term Memory sequence classification reveal that binary classification of gender, height, race and age reaches up to 89.6, 72.1, 63.3 and 57.6 % unweighted accuracy on beat level in unseen test data.

1. INTRODUCTION

Singer trait classification, that is, automatically recognizing meta data such as age and gender of the performing vocalist(s) in recorded music, is currently still an underresearched topic in music information retrieval, in contrast to the increasing efforts devoted to that area in paralinguistic speech processing. Speaker trait recognition is often used in dialog systems to improve service quality [1], yet another important area of application is forensics where it can deliver cues on the identities of unknown speakers [9]. Likewise, applications in music processing can be found in categorization and query of large databases with potentially unknown artists – that is, artists for whom not enough reliable training data is available for building singer identification models as, e. g., in [12]. Robustly extracting a variety of meta information can then allow the artist to be identified in a large collection of artist meta data, such as the Internet Movie Database (IMDB). In addition, exploiting gender information is known to be very useful for building models for other music information retrieval tasks such as automatic lyrics transcription [11].

Current research in speech processing suggests that the automatic determination of age in full realism is challenging even in clean, spoken language [16]. On the other hand, it is well known that age as well as body shape (height and weight) have acoustic correlates [4, 10] that can be exploited for automatic classification [13]; additionally, it has been shown that demographic traits including ethnicity can be derived from spoken language [6]. In comparison to speech, recognition of *singer* traits is expected to be an even more challenging task due to pitch variability, influence of voice training, and presence of multiple vocalists as well as instrumental accompaniment. Previous research dealt with gender identification of unseen artists in recorded popular music [17], which could be performed with over 90 % accuracy in full realism by extracting the leading voice through an extension of non-negative matrix factorization (NMF) [3].

Still, to our knowledge, few, if any, studies exist on recognition of other singer traits in music. Hence, we introduce three new dimensions to be investigated: age, height and race. Our annotation scheme is inspired by the TIMIT corpus commonly used in speech processing, which provides rich speaker trait information. As such, we adopt the term 'race' from the corpus' meta-information—though modern biology often neither classifies the homo sapiens sapiens by race nor sub-categories for collective differentiation in both physical and behavioral traits. While current molecular biologic and population genetic research argues that a systematic categorization may not suffice the enormous diversity and fluent differences between geographic population, it can be argued that when aiming at an end-user information retrieval application, a classification into illustrative, archetypal categories is desirable.

For evaluation of automatic singer-independent classification, we extended the UltraStar database [15] with detailed annotation of singer traits (Section 2). Furthermore, we improve extraction of the leading voice by filtering of drum accompaniment (Section 3). The classification by Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNN) is briefly outlined in Section 4. Comprehensive evaluation results are presented in Section 5 before conclusions are drawn in Section 6.

2. ULTRASTAR SINGER TRAITS DATABASE

Our experiments build on the UltraStar database proposed in [17] for singer-independent evaluation of vocalist gender recognition, containing 581 songs commonly used for the 'UltraStar' karaoke game, corresponding to over 37 h total play time. Note that using highly popular songs is no contradiction to the goal of recognizing unknown artists, but rather a requirement for establishment of solid ground truth. The database is split according to the first letter of the name of the performer into training (A, D, G, ...), development (B, E, H, ...) and test partitions (0-9, C, F, ...). The ground truth tempo is provided and lyrics are aligned to (quarter) beats. The annotation of the database was substantially extended beyond gender information: The identity of the singer(s) was determined at beat level wherever possible. This is particularly challenging in case of formations such as 'boy-' or 'girl-groups', in which case the 'singer diarization' (alignment of the singer identity to the music) was determined from publicly available music videos. Then, information on gender, height, birth year and race of the 516 distinct singers present in the database was collected and multiply verified from on-line textual and audiovisual knowledge sources, including IMDB, Wikipedia and YouTube. All annotation was performed by two male experts for popular music (24 and 28 years old).

In a multitude of cases, two or more singers are singing simultaneously. In [17], which only dealt with gender recognition, the case that male and female vocalists are singing in 'duet' was treated as a special case, where the corresponding beats were excluded from further analysis. To extend this paradigm to the now multi-dimensional annotation, we derived the following scheme: For nominal traits (gender and race), the beats were marked as 'unknown' unless all simultaneously present artists share the same attribute value. For continuous-valued traits (age and height), the average value was calculated, since in formations the individual artists' traits are usually similar. This procedure was also followed to treat performances of formations where an exact singer diarization could not be retrieved, by assuming presence of an 'average singer' throughout. In case that the desired attribute is missing for at least one of the performing vocalists, the corresponding beats were marked as 'unknown'.

The distribution of gender and race among the 516 singers are shown in Figures 1a and 1b. Age (Figure 1c) and height (Figure 1d) are shown as box-and-whisker plots where boxes range from the first to the third quartile and all values that exceed that range by more than 1.5 times the width of the box are considered outliers, depicted by circles. Unlike gender, height, and race, the age distribution can only be given on beat level since age is not well defined per artist (due to different recording dates) nor per song (due to potentially multiple singers per song). The continuous-valued attributes height and age were discretized to 'short' (s, $< 175\,cm$) and 'tall' (t, $\geq 175\,cm$), respectively 'young' (y, $<$ 30 years) and 'old' (o, $\geq$ 30 years); the thresholds are motivated by the medians of the traits (175 cm resp. 28 years) to avoid sparsity of either class. For race, the prototypical classes 'White', 'Black', 'Hispanic' and 'Asian' were annotated. The smaller classes 'Black', 'Hispanic' and 'Asian' were subsumed due to great sparsity of 'Hispanic' and 'Asian' singers: Our goal is to evaluate our system on all data for which a ground truth is available. 'Unknown' beats are excluded from further analysis. From the manual singer diarization and collection of singer meta data, the beat level annotation is generated automatically, resulting in the number of beats and according classification tasks shown in Table 1. To foster further research on the challenging topics introduced in this paper, the annotation (singer meta-data, voice alignments, song list with recording dates and partitioning) is made publicly available for research purposes at http://www.openaudio.eu.

# beats	**train**	**devel**	**test**	Σ
no voice (0)	90 076	75 741	48 948	214 765
		gender		
female (f)	32 308	23 071	9 739	65 118
male (m)	55 505	49 497	37 686	142 688
?	86	253	771	1 110
		race		
white (w)	67 525	62 003	40 479	170 007
b/h/a	16 378	9 465	7 136	32 979
?	3 996	1 353	581	5 930
		age		
young (y)	48 510	42 056	25 682	116 248
old (o)	34 074	24 596	18 712	77 382
?	5 315	6 169	3 802	15 286
		height		
short (s)	29 638	24 946	8 562	63 146
tall (t)	30 177	30 146	23 452	83 775
?	28 084	17 729	16 182	61 995
Σ	**177 975**	**148 562**	**97 144**	**423 681**

Table 1: Number of beats per trait, class and set (train / devel /test) in the UltraStar singer trait database. 'b/h/a': black / hispanic / asian. 'Unknown' (?) includes simultaneous performance of artists of different gender / race, as well as those with unknown ground truth.

3. MONAURAL SOURCE SEPARATION METHODS

A major part of our experiments is devoted to finding the optimal preprocessing by source separation for recognition of vocalist gender, age, height and race. To this end, we investigate harmonic enhancement as in [8, 17] and extraction of the leading voice as in [3], as well as a combination of both.

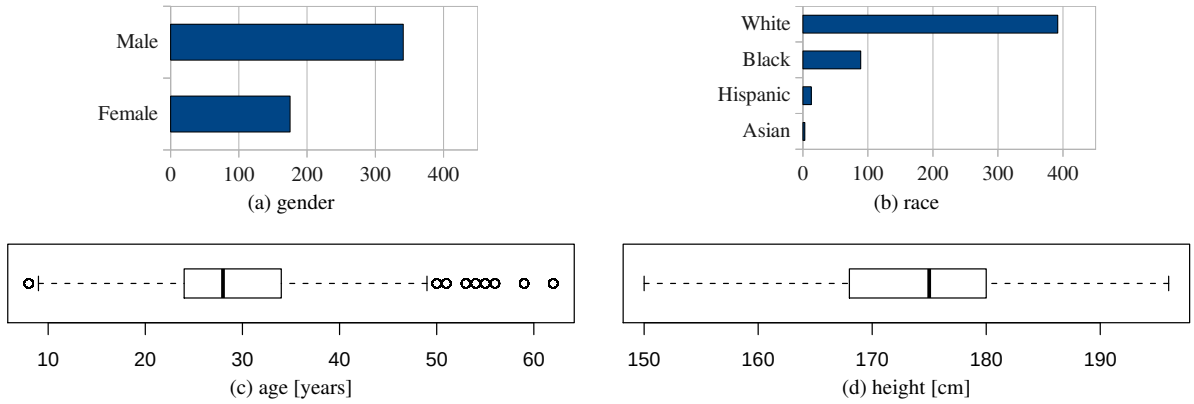

Figure 1: Distribution of gender, race, and height among 516 singers in the UltraStar Singer Trait Database. Distribution of age is shown on beat level, since it is dependent on recording date.

3.1 Enhancement of Harmonic Parts

Enhancement of harmonic parts is performed following [17]. It is based on a non-negative factorization of the magnitude spectrogram obtained by Short-Time Fourier transform (STFT) that is computed using a multiplicative update algorithm for NMF minimizing the Kullback-Leibler divergence. We then use a Support Vector Machine (SVM) to discriminate between components (spectra and their time-varying gains) corresponding to percussive or non-percussive signal parts. The classifier is trained on a manually labeled set of NMF components extracted from popular music as described in [15]. The features for discrimination of drum and harmonic components exactly correspond to those used in [15]. For straightforward reproducibility of our experiments, we used the default parameters of the publicly available [1] drum beat separation demonstrator of the source separation toolkit openBliSSART [18]: frame rate 30 ms, window size 60 ms, and 100 iterations. 50 NMF components are used; for 20 components thereof, the spectral shape $\mathbf{w}$ is pre-initialized from typical drum spectra delivered with the openBliSSART demonstrator. To allow the algorithm to use different sets of components for the individual sections of a song, chunking into frame-synchronous non-overlapping chunks is performed as in [17].

3.2 Leading Voice Separation

The second method used to facilitate singer trait identification is the leading voice separation approach described in [2, 3]. In this model, the STFT of the observed signal at each frame is expressed as the sum of STFTs of vocal and background music signals. These are estimated by an unsupervised approach: The voice STFT is modelled as product of source (periodic glottal pulse) and filter STFTs while no specific constraints are set for the background music signal because of its wide possible variability. The estimation of the various model parameters is then conducted by iterative approaches based on NMF techniques following a two step strategy. The first step provides an initial estimate of the parameters while the second step is a constrained re-estimation stage which refines the leading melody estimation and in particular limits sudden octave jumps that may remain after the first estimation stage. To ensure best reproducibility of our results, we used an open-source implementation [2] of the algorithm with default parameters. Chunking was applied as in [17].

3.3 Combined Source Separation Approaches

When the algorithm described in the last section is applied to popular music, it turns out that part of the drum track may remain after separation. Hence, for this study, we considered cascading of both separation techniques: harmonic enhancement after leading voice separation (LV-HE), and vice versa (HE-LV). Thereby time domain signals are synthesized inbetween the two separation stages, in order to be able to use different NMF parameterizations for both algorithms.

4. EXPERIMENTAL SETUP

4.1 Acoustic Features

The features exactly correspond to those used in [15] and were extracted for each beat using the open-source toolkit openSMILE [5]. We consider the short-time energy, zero-, and mean-crossing rate known to indicate vocal presence. In addition we extract values from the normalized autocorrelation sequence of the DFT coefficients, namely voicing probability, F-zero and harmonics-to-noise ratio (HNR). F-zero

[1] http://openblissart.github.com/openBliSSART

[2] Software available at http://www.durrieu.ch/phd/software.html

is the location of the highest peak of the autocorrelation sequence aside from the maximum at zero. HNR is computed by the value of this peak. Pitch and voice quality parameters have been successfully used in paralinguistic information assessment from speech [16]. We further calculate Mel frequency cepstral coefficients (MFCC) 0–12 and their respective first-order delta regression coefficients which are known to capture the characteristic qualities of individual voices for singer identification [12]. Thus, altogether we employ a set of 46 time-varying features. The employed configuration of the openSMILE toolkit is provided for further reproducibility at http://www.openaudio.eu.

4.2 Classification by BLSTM-RNN

As in [17], sequence classification with Bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks (RNNs) has been observed greatly superior to beat-wise static classification by SVMs or Hidden Naive Bayes on the vocalist gender recognition task (90.77 % beat level accuracy on original signals vs. 72.78 % resp. 76.17 %), we opt for this classifier for our study. BLSTM-RNNs unite the concept of bidirectional RNNs (BRNNs) with Long Short-Term Memory (LSTM) [7]. BRNNs use two separate hidden layers instead of one, both connected to the same input and output layers, of which the first processes the input sequence forwards and the second backwards. The network therefore always has access to the complete past and the future context in a symmetrical way. Consequently, it must have the complete input sequence at hand before it can be processed; however, this is not a restriction in the context of our application. In short, the LSTM concept allows the network to access potentially unlimited range of context, and to learn when to store, use, or discard information acquired from previous inputs or outputs. This makes (B)LSTM-RNNs useful for sequence classification tasks where the required amount of context is unknown a priori.

4.3 BLSTM Topology and Training

We trained individual BLSTM networks for each classification task. As in [17], the networks had one hidden layer with 80 LSTM memory cells for each direction. The size of the input layer was equal to the number of features (46), while the size of the output layer was equal to the number of classes to discriminate (2–3). Its output activations were restricted to the interval $[0; 1]$ and their sum was forced to unity by normalizing with the softmax function. Thus, the normalized outputs represent the posterior class probabilities. The songs in the test set were presented frame by frame (in correct temporal order) to the input layer, and each frame was assigned to the class with the highest probability as indicated by the output layer. For network training, supervised learning with early stopping was used as follows: We initialized the network weights randomly from a Gaussian distribution ($\mu = 0, \sigma = 0.1$). Then, each sequence (song) in the UltraStar training set was presented frame by frame to the network. To improve generalization, the order of the input sequences was determined randomly, and Gaussian noise ($\mu = 0, \sigma = 0.3$) was added to the input activations. The network weights were iteratively updated using resilient propagation [14]. To prevent over-fitting, the performance (in terms of classification error) on the validation set was evaluated after each training iteration (epoch). Once no improvement over 20 epochs had been observed, the training was stopped and the network with the best performance on the validation set was used as the final network. As the race recognition problem is particularly unbalanced, slight modifications were employed for the training procedure: A fixed number of 20 epochs was run to avoid overfitting to the validation set, and the standard deviation of the Gaussian noise on the input activations was increased to $\sigma = 0.9$.

5. RESULTS

Our primary measure for evaluating performance of automatic singer trait recognition is unweighted accuracy (UA)—i. e., the average recall of the classes—on beat level. Due to class imbalance (Table 1) it represents the discrimination power of the classifier more closely than 'conventional' weighted accuracy (WA) where recalls of the classes are weighted with their a-priori probabilities. Note that both, random guessing or always picking the majority class would achieve a UA of 33.33 % in ternary and 50.00 % in binary classification tasks.

5.1 Results on Beat Level

In order to highlight the difficulty of the evaluated singer trait recognition tasks in full realism, we first evaluated the BLSTM-RNN on the task to recognize the presence of a singer. It turns out that this can be done with over 75 % UA when using the leading voice extraction – note that this algorithm usually extracts the leading instrument when no voice is present, hence the task remains non-trivial. Best results on the 2-class gender recognition task are obtained by the proposed combination of source separation algorithms (LV-HE, 89.61 % UA) while in the 3-class task, best UA is achieved by the LV algorithm alone (69.29 % UA). Notably, this is higher than it would be expected if accuraries of voice activity and 2-class gender recognition were independent. 2-class recognition of race delivers up to 63.30 % UA when including HE preprocessing, while LV alone downgrades performance compared to the original. Furthermore, we observe that *height* recognition can be robustly performed at up to 72.07 % UA when using HE-LV preprocessing, which boosts the UA by over 7 % absolute compared to no pre-

[%]		Preprocessing –		HE		LV		LV-HE		HE-LV	
task	classes	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA
voice	0/1	74.55	74.50	73.82	73.84	**75.77**	75.81	75.40	75.41	75.09	75.11
gender	0/m/f	63.75	68.54	65.65	68.91	**69.29**	71.31	67.90	70.41	68.52	70.44
	m/f	86.67	91.09	88.45	91.91	86.93	91.12	**89.61**	93.60	87.76	92.50
race	0/w/b+h+a	48.17	63.84	47.46	63.02	**49.37**	65.46	49.23	63.63	48.40	63.77
	w/b+h+a	60.44	65.82	**63.30**	76.98	55.05	76.18	62.57	78.67	62.78	75.16
age	0/y/o	51.02	57.61	50.00	57.14	**53.50**	59.85	51.26	58.86	50.01	57.72
	y/o	55.30	55.60	**57.55**	56.56	53.93	53.63	55.97	54.89	54.69	54.17
height	0/s/t	53.94	66.79	52.35	66.57	58.15	69.30	57.67	68.41	**58.91**	69.53
	s/t	64.70	72.73	62.31	70.67	66.54	73.00	69.65	77.49	**72.07**	78.26

Table 2: Beat-wise BLSTM-RNN classification of UltraStar test set on 2- and 3-class tasks. Preprocessing: HE = harmonic enhancement (Section 3.1); LV = leading voice extraction (Section 3.2); LV-HE: HE after LV; HE-LV: LV after HE.

[%]		Preprocessing –		HE		LV		LV-HE		HE-LV	
task	vote on	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA
gender	0/m/f	80.9	87.0	81.7	85.6	87.7	90.9	**91.3**	92.4	87.7	90.9
	m/f	86.9	90.1	89.0	90.9	87.7	90.9	**89.6**	93.9	89.6	93.9
race	0/w/b+h+a	49.8	78.8	53.5	79.7	51.0	78.2	**54.0**	75.2	48.9	72.2
	w/b+h+a	52.8	59.8	62.6	75.9	54.7	73.7	**64.4**	78.9	61.7	74.4
age	0/y/o	55.2	54.5	54.6	54.1	56.0	54.1	**56.9**	57.4	50.9	51.6
	y/o	54.5	54.5	57.0	55.7	52.2	51.6	53.4	52.5	**58.9**	58.2

Table 3: Song-wise BLSTM-RNN predictions on UltraStar test set by beat-wise majority vote. Vote among 3-class tasks (ignoring beats not classified as 0) or 2-class tasks. Height is not included due to the low number of songs (88) with known ground truth. Preprocessing as in Table 2.

processing. Finally, up to 57.55 % UA are achieved in *age* recognition when using HE; while this is clearly below typical results on spoken language, it is significantly above chance level (50 % UA) according to a z-test ($p < .001$).

5.2 Results on Song Level

As a performance estimate for 'tagging' entire songs, we calculated for each scenario the accuracies of majority vote on beat level compared with the most frequent ground truth class on beat level. Note that such measurements are more heuristic in nature, since a song level ground truth cannot always be established due to typical phenomena in real-life music such as alternating male / female vocalists. To briefly comment on the results, song level gender can be recognized with up to 91.3 % UA, race with 64.4 % UA and age with 58.9 % UA. For gender, estimation from the vote on *all* (not only voiced) beats seems to be even more robust than votes on the 2-class beat level task. LV-HE preprocessing delivers overall best results.

5.3 Discussion and Outlook

For *race*, an interdependency with genre could be assumed; however, the fact that source separation generally improves the result over the original music suggests that genre is not the primary information learned by the classifier. Furthermore, genres typically associated with non-white singers such as hip hop are very sparsely represented in the UltraStar database, which is originally intended for karaoke applications. Still, the very robust recognition of *height* is clearly correlated with robust gender identification, as tall female singers are sparse in the considered data set.

Compared to 'usual' results obtained on spoken language, accuracies of *age* recognition are rather low; the task seems to be especially challenging on the considered type of 'chart' popular music with a prevalence of singers in their twenties. At least, when using gender-dependent models for age, 61.63 % UA could be achieved for males; for females there is not enough training data.

A promising direction for further research may be to investigate different units of analysis, such as longer-term statistical functionals that are commonly used in paralinguistic information retrieval from speech [16], instead of recogni-

tion at the beat level. Still, this is not fully straightforward due to the feature variation, especially for pitch, which will necessitate methods for robust pitch estimation and transformation.

6. CONCLUSIONS

Inspired by previous successful studies on vocalist gender recognition, we introduced fully automatic assessment of new paralinguistic traits (age, height and race) in a large collection of recorded popular music. While we could also improve gender recognition close to perfection even on beat level (up to 93.60 % WA on unseen test data), foremost we have shown feasibility of race and height classification in full realism. Even in chart music with a prevalence of singers from 20–30 years, age recognition could be performed significantly above chance level; still, when aiming at real-life applications new directions in research must be taken.

Future work should primarily focus on more variation in data (particularly concerning age and race) by not only including chart music, but also jazz and non-Western music. Furthermore, we will investigate multi-task learning to exploit singer trait interdependencies in learning.

7. ACKNOWLEGMENT

The research leading to these results has been partly funded by the German Research Foundation through grant no. SCHU 2580/2-1. The authors would like to thank Gaël Richard and Jean-Louis Durrieu for their highly valuable contributions.

8. REFERENCES

[1] F. Burkhardt, R. Huber, and A. Batliner. Application of speaker classification in human machine dialog systems. In Christian Müller, editor, *Speaker Classification I: Fundamentals, Features, and Methods*, pages 174–179. Springer, 2007.

[2] J.-L. Durrieu, G. Richard, and B. David. An iterative approach to monaural musical mixture de-soloing. In *Proc. of ICASSP*, pages 105–108, Taipei, Taiwan, 2009.

[3] J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575, 2010.

[4] S. Evans, N. Neave, and D. Wakelin. Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology*, 72(2):160–163, 2006.

[5] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – the Munich versatile and fast open-source audio feature extractor. In *Proc. of ACM Multimedia*, pages 1459–1462, Florence, Italy, October 2010. ACM.

[6] D. Gillick. Can conversational word usage be used to predict speaker demographics? In *Proc. of Interspeech*, pages 1381–1384, Makuhari, Japan, 2010.

[7] A. Graves. *Supervised sequence labelling with recurrent neural networks*. PhD thesis, Technische Universität München, 2008.

[8] M. Helén and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. of EUSIPCO*, Antalya, Turkey, 2005.

[9] M. Jessen. Speaker classification in forensic phonetics and acoustics. In C. Müller, editor, *Speaker Classification I*, volume 4343, pages 180–204. Springer Berlin / Heidelberg, 2007.

[10] R. M. Krauss, R. Freyberg, and E. Morsella. Inferring speakers physical attributes from their voices. *Journal of Experimental Social Psychology*, 38(6):618–625, 2002.

[11] A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009. Article ID 546047.

[12] A. Mesaros, T. Virtanen, and A. Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proc. of ISMIR*, pages 375–378, 2007.

[13] I. Mporas and T. Ganchev. Estimation of unknown speakers height from speech. *International Journal of Speech Technology*, 12(4):149–160, 2009.

[14] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In *Proc. of IEEE International Conference on Neural Networks*, pages 586–591, 1993.

[15] B. Schuller, C. Kozielski, F. Weninger, F. Eyben, and G. Rigoll. Vocalist gender recognition in recorded popular music. In *Proc. of ISMIR*, pages 613–618, Utrecht, Netherlands, August 2010.

[16] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The INTERSPEECH 2010 Paralinguistic Challenge. In *Proc. of INTERSPEECH*, pages 2794–2797, Makuhari, Japan, September 2010. ISCA.

[17] F. Weninger, J.-L. Durrieu, F. Eyben, G. Richard, and B. Schuller. Combining Monoaural Source Separation With Long Short-Term Memory for Increased Robustness in Vocalist Gender Recognition. In *Proc. of ICASSP*, Prague, Czech Republic, 2011.

[18] F. Weninger, A. Lehmann, and B. Schuller. openBliSSART: Design and Evaluation of a Research Toolkit for Blind Source Separation in Audio Recognition Tasks. In *Proc. of ICASSP*, Prague, Czech Republic, 2011.