COMBINING MONAURAL SOURCE SEPARATION WITH LONG SHORT-TERM MEMORY FOR INCREASED ROBUSTNESS IN VOCALIST GENDER RECOGNITION

Felix Weninger[1]*, Jean-Louis Durrieu*[2]*, Florian Eyben*[1]*, Gaël Richard*[3]*, and Björn Schuller*[1]

[1]Institute for Human-Machine Communication, Technische Universität München, D-80290 München
[2]Signal Processing Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne
[3]Institut Télécom, Télécom ParisTech, LTCI-CNRS, F-75014 Paris
weninger@tum.de

ABSTRACT

We present a novel and unique combination of algorithms to detect the gender of the leading vocalist in recorded popular music. Building on our previous successful approach that enhanced the harmonic parts by means of Non-Negative Matrix Factorization (NMF) for increased accuracy, we integrate on the one hand a new source separation algorithm specifically tailored to extracting the leading voice from monaural recordings. On the other hand, we introduce Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNNs) as context-sensitive classifiers for this scenario, which have lately led to great success in Music Information Retrieval tasks. Through a combination of leading voice separation and BLSTM networks, as opposed to a baseline approach using Hidden Naive Bayes on the original recordings, the accuracy of simultaneous detection of vocal presence and vocalist gender on beat level is improved by up to 10 % absolute. Furthermore, using this technique we achieve 91.6 % accuracy in determining the gender of the predominant vocalist on song level, which is 4 % absolute above our previous best result.

Index Terms— Long Short-Term Memory, Non-Negative Matrix Factorization, Music Information Retrieval

1. INTRODUCTION

Vocalist gender recognition, that is determination of the gender of the main vocalist(s) in recorded (popular) music, is a task that has not yet been broadly addressed in the field of Music Information Retrieval (MIR), as opposed to spoken language processing [1]. On the other hand, it is considerably challenging when performed on contemporary popular music due to the variety of singing styles and the large spectral overlap with the instrumental accompaniment. As in speech processing, knowing the gender of the lead performing artist may be used to select gender-adapted models for lyrics transcription – such as in [2] – or other MIR tasks. Additionally, it can be useful as a feature for organizing and querying music collections, or for recommendation systems in on-line stores, which motivated the introduction of vocalist gender recognition in [3].

Building on that study, the first major contribution of this paper is to compare different solutions to reduce the accompaniment in order to more reliably identify the singing voice, including both leading voice separation and enhancement of harmonic parts by drum beat separation. The former appears to be naturally suited to the recognition task at hand, and we have recently introduced a technique incorporating Non-Negative Matrix Factorization (NMF) on a source / filter model and Viterbi-based melody smoothing, which robustly identifies the pitch of the leading voice [4], then separates it from the signal [5]. On the other hand, in [3] we used enhancement of harmonic parts to great success, which can be performed very robustly by discrimination of the drum and harmonic signal parts extracted by Non-Negative Matrix Factorization, based on spectral and temporal characteristics [6].

Moreover, we depart from the static beatwise classification in [3] which completely ignores context. As it can be argued that context is vital in determining the gender of the lead singer, which is expected to change rather slowly over time, we introduce Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNNs) as a context-sensitive sequence classifier. Their most prominent advantage over other sequence classifiers is that they automatically learn the required amount of context [7]. Notably, we have used BLSTM-RNNs for our onset detector which performed best among all approaches evaluated in the 2010 Music Information Retrieval EXchange (MIREX) challenge [8].

The remainder of this paper is structured as follows: first, we summarize our source separation methods for enhancement of harmonic parts and leading voice separation in Sec. 2. Next, we briefly present the basic concept of BLSTM networks in Sec. 3 before turning to a detailed description of our experimental setup and classification procedures in Sec. 4. Results are interpreted, and conclusions are drawn in Sec. 5. To increase clarity of the following section, we introduce the following notations: for a matrix $\mathbf{A}$, the notation $[\mathbf{A}]_{i,:}$ – resembling Matlab syntax – denotes the i-th row of $\mathbf{A}$ (as a row vector), and we analogously define $[\mathbf{A}]_{:,j}$ for the j-th column of $\mathbf{A}$ (as a column vector). We write $\mathbf{A} \otimes \mathbf{B}$ for the elementwise product of matrices $\mathbf{A}$ and $\mathbf{B}$; division of matrices is always to be understood as elementwise.

2. SOURCE SEPARATION METHODS

2.1. Enhancement of Harmonic Parts

As a first method to improve gender identification from the vocal parts, we chose the enhancement of harmonic parts as in our previous study [3]. It is based on a non-negative factorization of the magnitude spectrogram $|\mathbf{X}|$ obtained by short-time transform (STFT):

$$|\mathbf{X}| = \mathbf{W}\mathbf{H},$$

that is computed using a multiplicative update algorithm for NMF which minimizes the β-divergence between $|\mathbf{X}|$ and $\mathbf{W}\mathbf{H}$, for $\beta = 1$

This work was supported by the German Research Foundation (DFG) under the grant no. SCHU 2508/2-1 ("Non-Negative Matrix Factorization for Robust Feature Extraction in Speech Processing"), and partly supported by the OSEO-Quaero program.

(Kullback-Leibler divergence). We then use a Support Vector Machine (SVM) classifier to discriminate between drum and harmonic components, i. e. pairs $(\mathbf{w}^{(j)}, \mathbf{h}^{(j)})$ of spectra $\mathbf{w}^{(j)} := [\mathbf{W}]_{:,j}$ along with their time-varying gains $\mathbf{h}^{(j)} := [\mathbf{H}]_{j,:}$ that model percussive or non-percussive signal parts. The classifier is trained on a set of NMF components extracted from popular music that were manually labeled as 'drum' or 'harmonic', as described in [3]. The features for discrimination of drum and harmonic components exactly correspond to those used in [3]. After classification, we estimate the magnitude spectrogram $|\widehat{\mathbf{X}}|_{\text{harm}}$ of the harmonic signal parts as follows: defining

$$J_{\text{harm}} = \{j : (\mathbf{w}^{(j)}, \mathbf{h}^{(j)}) \text{ classified as harmonic}\},$$

$$|\widehat{\mathbf{X}}|_{\text{harm}} = |\mathbf{X}| \otimes \sum_{j \in J_{\text{harm}}} \frac{\mathbf{w}^{(j)}\mathbf{h}^{(j)}}{\mathbf{WH}}. \quad (1)$$

Finally, we obtain a 'harmonic' time signal from the inverse STFT of $|\widehat{\mathbf{X}}|_{\text{harm}}$ and the phase matrix of the original signal, thereby windowing each time frame. Note that this procedure goes beyond the one used for our previous study [3]: due to the Wiener filtering in (1), it is guaranteed that the estimated spectrograms of the harmonic and non-harmonic signal parts sum to the original spectrogram. Therefore, no information is lost due to the factorization and reconstruction. For straightforward reproducibility of our experiments, we used the default parameters of the drum beat separation demonstrator of our toolkit openBliSSART[1]: frame rate 30 ms, window size 60 ms, and 100 iterations. Although the algorithm can efficiently process songs as a whole, we found it beneficial to divide the songs into non-overlapping chunks of 19.98 s length (synchronous to the frame rate). This allows the algorithm to use different sets of components for the individual sections of a song.

2.2. Leading Voice Separation

The second method used to facilitate gender identification is the leading voice separation approach described in [4, 5]. In this model, the STFT of the observed signal at frame n, denoted $[\mathbf{X}]_{:,n}$, is expressed as the sum of two components as $[\mathbf{X}]_{:,n} = [\mathbf{V}]_{:,n} + [\mathbf{M}]_{:,n}$, where $[\mathbf{V}]_{:,n}$ and $[\mathbf{M}]_{:,n}$ are respectively the STFTs of the leading voice and background musical signals. Furthermore, $[\mathbf{V}]_{:,n}$ and $[\mathbf{M}]_{:,n}$ are assumed to be center proper complex Gaussian variables[2]:

$$[\mathbf{V}]_{:,n} \sim \mathcal{N}_c(0, \text{diag}(\sigma^2_{[\mathbf{V}]_{:,n}})), \quad (2)$$

$$[\mathbf{M}]_{:,n} \sim \mathcal{N}_c(0, \text{diag}(\sigma^2_{[\mathbf{M}]_{:,n}})), \quad (3)$$

where $\sigma^2_{[\mathbf{V}]_{:,n}}$ (resp. $\sigma^2_{[\mathbf{M}]_{:,n}}$) is the power spectral density (PSD) of the leading voice (resp. of the background music) at frame n. Following an independence assumption between the two components, the STFT of the observed signal is also a proper Gaussian vector:

$$[\mathbf{X}]_{:,n} \sim \mathcal{N}_c(0, \text{diag}(\sigma^2_{[\mathbf{V}]_{:,n}} + \sigma^2_{[\mathbf{M}]_{:,n}})). \quad (4)$$

Extracting the main melody then consists in estimating $\sigma^2_{[\mathbf{V}]_{:,n}}$ and $\sigma^2_{[\mathbf{M}]_{:,n}}$ for each signal frame n. Here, the approach is entirely unsupervised (i. e. no learning step is involved) and therefore relies on specific constraints for the voice signal. More precisely, the voice signal is assumed to follow a source / filter production model where the source is a periodic signal (referring to the periodic glottal pulse of the singing voice). 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. Once the PSD $\sigma^2_{[\mathbf{V}]_{:,n}}$ and $\sigma^2_{[\mathbf{M}]_{:,n}}$ of both signals are obtained, the separated singing voice signal is obtained by Wiener filtering for each frame :

$$\widehat{[\mathbf{V}]}_{:,n} = \frac{\sigma^2_{[\mathbf{V}]_{:,n}}}{\sigma^2_{[\mathbf{V}]_{:,n}} + \sigma^2_{[\mathbf{M}]_{:,n}}} [\mathbf{X}]_{:,n}. \quad (5)$$

To ensure best reproducibility of our results, we used our open-source implementation[3] of the algorithm with default parameters (frame rate 256 samples, window size 2 048 samples, 50 iterations for each of the first and second separation stage). To cope with the memory requirements of the algorithm, it was applied to frame-synchronous chunks of 881 664 samples ($\approx$ 20 s at 44.1 kHz sample rate).

3. BIDIRECTIONAL LONG SHORT-TERM MEMORY RECURRENT NEURAL NETWORKS

A key part of the study presented in this paper is to evaluate the performance of BLSTM-RNNs as context-sensitive sequence classifiers on the vocalist gender recognition task. BLSTM-RNNs unite the concept of bidirectional RNNs (BRNNs) with Long Short-Term Memory (LSTM). Here we will only briefly describe the theory underlying BLSTM networks to motivate their use in this study. For a detailed discussion of the network architecture, we refer to [7].

RNNs consist of one input, one output and one or more hidden layer(s). In contrast to basic feedforward neural networks, cyclic connections theoretically allow the network to map from the entire history of previous inputs to an output. The recurrent connections form a kind of memory, which allows input values to persist in the hidden layer(s) and influence the network output in the future. 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.

Although (B)RNNs have access to past (and future) information, the range of context is limited to a few frames due to the vanishing gradient problem: the influence of an input value decays or blows up exponentially over time. To overcome this deficiency, the LSTM concept was introduced. In an LSTM hidden layer, the nonlinear units are extended to LSTM memory blocks. Each block contains one or more linear memory units, whose internal state is maintained by a recurrent connection with constant weight 1.0, enabling the unit to store information over arbitrary periods of time. The input, output, and internal state of the memory units are controlled by multiplicative gate units, which - in computer memory terminology - correspond to write, read, and reset operations. The gates are connected to the input layer as well as recurrently to the output layer. During network training, the weights for all connections, including the gate units, are optimized such that the network automatically learns when to store,

[1] Software available at http://www.openaudio.eu

[2] A complex proper Gaussian random variable is a complex random variable whose real part and imaginary part are independent and follow a (real) Gaussian distribution, with the same parameters: mean equal to 0 and identical variance (co-variance matrix in the multi-variate case).

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

# beats	train	devel	test	Σ
no voice	87 592	74 174	45 525	207 291
female	33 194	21 949	10 576	65 719
male	52 178	48 842	34 718	135 738
duet	370	202	1 130	1 702
Σ	**173 334**	**145 167**	**91 949**	**410 450**

Table 1: Number of beats per class and set in the UltraStar database.

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. They have been successfully used for a great variety of applications including handwriting recognition [7], automatic speech recognition [9], and note onset detection [8], often outperforming more traditional sequence classifiers such as Hidden Markov Models.

4. EXPERIMENTS

4.1. UltraStar Database

To evaluate the combination of source separation techniques and BLSTM-RNNs and enforce comparability of results, we chose the UltraStar song database introduced in [3], which is annotated on beat as well as song level. While the songwise, singer-independent subdivision into training, development, and test set was exactly preserved, we refined our annotation, i. e. the ground truth of the vocalist gender, to better reflect the real-world nature of the database where the gender may change several times throughout a song. Instead of assigning the predominant vocalist gender per song and propagating this decision to the beat level, as done in [3], we first labeled each beat as 'no voice', 'female voice', 'male voice', or as 'duet' if a male and female singer were present simultaneously. Then, we assigned the labels 'male' or 'female' on song level based on the songwise majority vote of the ground truths on the beats labeled as 'male' or 'female'. Furthermore, the alignment of the lyrics as well as the ground truth tempo was slightly corrected. The number of beats per class is shown in Tab. 1.

4.2. Baseline Classifiers

As in our previous study [3], the baseline classification performance is measured using SVM with polynomial kernel and Hidden Naive Bayes (HNB) [10]. SVMs were trained using Sequential Minimal Optimization (SMO). HNB was applied in a feature space that was discretized using Kononenko's Minimum Description Length (MDL) criterion. The features exactly correspond to those used in [3] and were extracted using the open-source toolkit openSMILE [11]. In line with our strategy to use open-source software for easy reproducibility, classification was performed using the Weka toolkit. In comparison to [3], the classification process was optimized in two respects: first, instead of downsampling the training material, we applied upsampling by copying the instances of the minority classes to achieve a roughly uniform class distribution among the union of training and development set. Second, we found it beneficial to lower the complexity parameter used for SVM training from 1.0 to 0.1.

4.3. BLSTM Topology and Training

The BLSTM 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, while the size of the output layer was equal to the number of classes to discriminate. 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. 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.

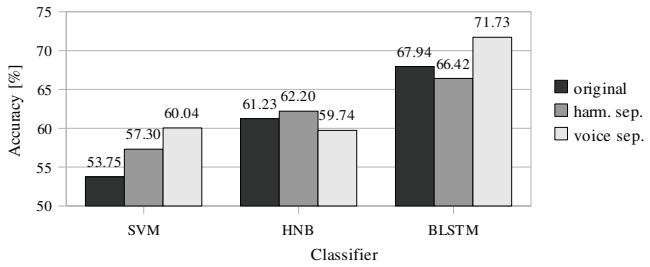
4.4. Beat and Song Level Classification

Each combination of classifier and source separation algorithm was evaluated on four different tasks. Thereby the beats labeled as 'duet' were excluded due to the very small number of instances. First, each beat in the test set had to be classified as 'male', 'female', or 'no voice' (3-class beatwise decision). Second, we restricted the training and testing material for the classifier to the beats where a voice was present according to the annotation, leading to a 2-class beatwise decision task. Finally, from the results of both these classification tasks, a song level decision can be performed by taking the majority vote of the beatwise decisions, considering only the beats classified as 'male' or 'female'.

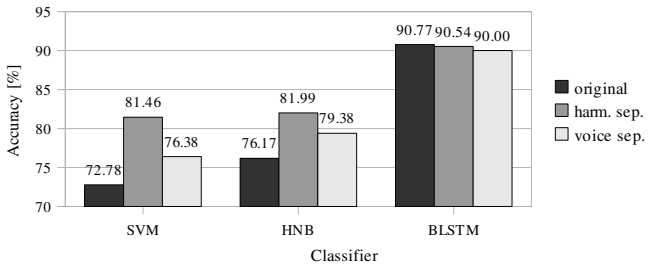
4.5. Results

Fig. 1 shows the classification accuracy achieved in the beatwise 3- and 2-class tasks. Comparing the classifiers, it can be clearly seen that the BLSTM outperforms static classifiers: even for the original audio, the BLSTM accuracy is 6.7 % absolute above the HNB accuracy. In particular, the combination of BLSTM with leading voice separation seems to be especially powerful on the 3-class task: here, the voice separation yields a gain in accuracy of almost 4 % absolute. Both the aforementioned improvements are significant at the 0.1 % level according to a one-tailed t-test. For the 2-class task, the BLSTM achieves excellent results independent of the preprocessing. While the SVM and HNB seem both inferior to the BLSTM in this task, it is notable that for these types of classifiers the harmonic enhancement conveys a larger gain than the leading voice separation (6 % vs. 3 % absolute for HNB).

Additionally, we present the results on song level, i. e. the weighted and unweighted average recall (WAR / UAR) of songs with a predominant female or male artist, in Tab. 2. Interestingly, the results obtained from the 3-class classifier on beat level seem to be more robust on song level than those from the 2-class classifier. Next, concerning different source separation methods as preprocessing, there is no clear picture: while the leading voice separation performs considerably better in the voting based on the 2-class task, achieving the overall best UAR, harmonic enhancement does so for the voting based on the 3-class task. Yet, since all these differences fail to be significant on the 5 % level, we conclude that the quality of the song level decision is highly robust against the preprocessing method.



(a) 3-class beatwise classification



(b) 2-class beatwise classification

Fig. 1: Accuracy in beatwise classification: 3-class task (no voice / female / male) and 2-class task (female / male), using SVM, HNB, and BLSTM classifiers. "voice sep." and "harm. sep." indicate preprocessing according to Secs. 2.1 and 2.2, respectively.

[%]	original		harm. sep.		voice sep.	
task	**WAR**	**UAR**	**WAR**	**UAR**	**WAR**	**UAR**
3-class	92.4	88.3	**93.1**	**88.8**	90.8	87.3
2-class	**91.6**	87.8	90.8	84.0	**91.6**	**88.9**

Table 2: Accuracy (weighted average recall, WAR) and unweighted average recall (UAR) of the song level decision based on 3-class and 2-class beatwise classification, using a BLSTM classifier. "voice sep." and "harm. sep." indicate preprocessing according to Secs. 2.1 and 2.2, respectively.

5. CONCLUSIONS

We have shown that our novel BLSTM-NMF approach yields a significant performance gain in the task to simultaneously detect voiced beats and to tell apart male and female vocalists. In fact, the apparently easier 2-class task to discriminate between male and female voiced beats can be performed more robustly, especially using the BLSTM classifier. However, our results indicate that the gender discrimination on song level is equally robust when using the results from the 3-class task. Most importantly, it can be argued that this task is the more realistic one, since it reflects the most common situation where the alignment of lyrics in the songs is unknown.

As a consequence, the performance gain on beat level by our novel technique to combine BLSTM and leading voice separation is especially valuable for a real-world application scenario. Furthermore, this performance gain is particularly interesting since the leading voice separation method is entirely unsupervised and further improvements can therefore be expected by integrating a training or adaptation stage. On the other hand, the beat level decisions of 3-class BLSTM classifier deliver an accuracy of up to 93.1 % on song level, which is almost 6 % absolute above our previous best result [3].

Building on these promising achievements, we will consider a source separation procedure that integrates leading voice separation and enhancement of harmonic components. Additionally, we will investigate how to best deal with simultaneous performance of artists. To this end, we will have to overcome limitations of the source separation process – the leading voice separation algorithm can only handle one main melody – as well as the classification. For example, we might consider a BLSTM regression (instead of classification) on voice activity for male / female artists. On the other hand, we might try to train with artificial mixtures to increase training material especially for the 'duet' class.

6. ACKNOWLEDGMENT

The authors would like to thank the student assistants Pascal Staudt and Christoph Kozielski for their highly valuable contributions.

7. REFERENCES

[1] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. of LREC*, Genoa, Italy, 2006.

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

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

[4] 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*, vol. 18, no. 3, pp. 564–575, 2010.

[5] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[6] 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.

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

[8] S. Böck, F. Eyben, and B. Schuller, "Onset detection with bidirectional long-short term memory neural networks," Music Information Retrieval EXchange (MIREX) 2010, available from http://www.music-ir.org/mirex/abstracts/2010/BES1.pdf.

[9] M. Wöllmer, F. Eyben, B. Schuller, Y. Sun, T. Moosmayr, and G. Rigoll, "Robust in-car spelling recognition - a tandem BLSTM-HMM approach," in *Proc. of Interspeech*, Brighton, UK, 2009.

[10] H. Zhang, L. Jiang, and J. Su, "Hidden Naive Bayes," in *Proc. of AAAI*, Pittsburgh, PA, USA, 2005, pp. 919–924.

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