COMBINING HMM-BASED MELODY EXTRACTION AND NMF-BASED SOFT MASKING FOR SEPARATING VOICE AND ACCOMPANIMENT FROM MONAURAL AUDIO

Yun Wang, Zhijian Ou

Department of Electronic Engineering, Tsinghua University, Beijing, China
Emails: maigoakisame@yahoo.com.cn, ozj@tsinghua.edu.cn

ABSTRACT

Modern monaural voice and accompaniment separation systems usually consist of two main modules: melody extraction and time-frequency masking. A main distinction between different separation systems lies in what approaches are used for the two modules. Popular techniques for melody extraction include hidden Markov models (HMMs) and non-negative matrix factorization (NMF), and masking includes hard and soft masking. This paper investigates the flaw of NMF-based melody extraction, and proposes the combination of HMM-based melody extraction (equipped with a newly-defined feature) and NMF-based soft masking. Evaluations on two publicly available databases show that the proposed system reaches state-of-the-art performance and outperforms several other combinations.

Index Terms— Monaural sound separation, melody extraction, soft masking, HMMs, NMF

1. INTRODUCTION

Modern monaural voice and accompaniment separation systems usually consist of two main modules: melody extraction and time-frequency masking. Popular techniques for melody extraction include hidden Markov models (HMMs), non-negative matrix factorization (NMF), and so on. Time-frequency masking decomposes an audio into time-frequency (T-F) units, then assigns each unit to the sound sources according to a certain proportion (a "mask"), and finally resynthesizes each sound source. Masking methods can be divided into "hard masking", where the mask consists of only zeros and ones, and "soft masking", where the elements in the mask can range continuously from 0 to 1.

A main distinction between different separation systems lies in what approaches are used for the two modules of melody extraction and masking. The computational auditory scene analysis (CASA) approach proposed by Wang *et al.*[1] employs HMM-based melody extraction and hard masking. In [2], Hsu *et al.*combine the melody extraction algorithm of Dressler [3], which makes use of neither HMM nor NMF, and the hard masking of Wang *et al.*, and in [4] they propose their own HMM-based melody extraction method. (For simplicity, we shall refer to the two systems of Hsu *et al.*as the H1 and H2 systems respectively.) Virtanen *et al.*[5] use the HMM-based melody extraction of Klapuri [6] and NMF-based soft masking. Durrieu *et al.*[7] perform melody extraction with NMF, and soft masking with both NMF and Wiener filtering.

Durrieu's system presents a promising approach. Regarding the masking step, it has been shown in [8] that Wiener filtering, which is closely related to the ideal ratio masking, performs consistently better than hard masking in terms of signal-to-noise ratio (SNR). However, the NMF-based melody extraction module tends to estimate the pitch as one octave higher, and results in octave errors. This undesirable effect is mentioned and ad hoc compensated in [7], but the underlying problem is unclear.

This paper looks into the reason why the octave errors are produced, and reveals the flaw of NMF-based melody extraction. Based on these discoveries, we propose a new monaural voice and accompaniment separation system, which combines HMM-based melody extraction and NMF-based soft masking. Our HMM-based melody extraction method draws on the H2 system [4] and Klapuri's work [6], and is equipped with a newly-defined feature. Evaluations on two publicly available databases show that our system performs better than both the H1 system and Durrieu's separation system, and that the HMM-based melody extraction runs significantly faster than the NMF-based approach.

This paper is organized as follows: In Section 2, we review Durrieu's algorithm and investigate its flaw. In Section 3, we describe our new separation system. Evaluation results are provided in Section 4, comparing our system with the H1, H2 and Durrieu's systems. Finally, the conclusions are made in Section 5.

2. REVIEW OF NMF-BASED MELODY EXTRACTION AND SOFT MASKING

Here we recapitulate the NMF-based melody extraction and soft masking initially proposed in [9]. It is slightly improved by introducing a smoothing matrix for the vocal track [10], which is included in our system but will be omitted in the following description.

The monaural signal $x(\tau)$ is considered to be a mixture of the voice $v(\tau)$ and the accompaniment $m(\tau)$. Its short-time Fourier transform spectrogram is denoted as a matrix $\boldsymbol{X}$. Each element X_{ft} at frequency bin f and frame t is assumed to obey a circular complex Gaussian distribution with zero mean and variance D_{ft}:

$$p(X_{ft}|D_{ft}) = \frac{1}{\pi D_{ft}} \exp\left(-\frac{|X_{ft}|^2}{D_{ft}}\right) \tag{1}$$

The variances of each element form a matrix $\boldsymbol{D}$, which is called the power spectrogram of $x(\tau)$.

Assuming that the voice and the accompaniment are independent, and further applying NMF, we can decompose the power spectrogram of the mixed signal $\boldsymbol{D}$ as follows:

$$\boldsymbol{D} = \underbrace{(\boldsymbol{B}_F\boldsymbol{A}_F).*(\boldsymbol{B}_K\boldsymbol{A}_K)}_{\boldsymbol{D}_V} + \underbrace{(\boldsymbol{B}_M\boldsymbol{A}_M)}_{\boldsymbol{D}_M} \tag{2}$$

Here ".*" denotes element-wise multiplication. The three parentheses on the right hand side stand for the power spectrograms of the

This work is supported by National Natural Science Foundation of China (61075020). The authors would like to thank J.-L. Durrieu for providing his separation program for evaluation.

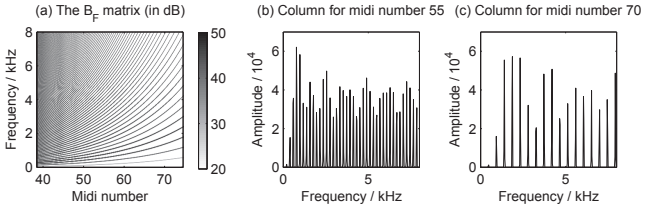

Fig. 1. The $\boldsymbol{B}_F$ matrix used in the proposed system

voice glottal excitation (with subscript F), the vocal tract response (with subscript K), and the accompaniment (with subscript M), respectively. In each parenthesis, the columns of the $\boldsymbol{B}$ matrix can be regarded as power spectrum bases, while the $\boldsymbol{A}$ matrix can be treated as linear combination coefficients of the bases.

The matrix of glottal excitation power spectrum bases, $\boldsymbol{B}_F$, is fixed and generated from the KLGLOTT88 model [11]. Its columns contain the power spectra of the glottal excitations at different f0's of interest. These f0's are chosen to be equally spaced on the midi number scale. The relationship between the midi number n and the frequency f (in Hertz) is:

$$n(f) = 69 + 12\log_2(f/440) \tag{3}$$

$$f(n) = 440 \times 2^{(n-69)/12} \tag{4}$$

Fig. 1 shows the $\boldsymbol{B}_F$ matrix used in our system. We choose the midi numbers from 38.5 to 74.5 with a step of 0.1.

2.1. NMF-based melody extraction

First, the five unknown matrices $\Theta = \{\boldsymbol{A}_F, \boldsymbol{B}_K, \boldsymbol{A}_K, \boldsymbol{B}_M, \boldsymbol{A}_M\}$ are estimated by maximizing the likelihood of the observed spectrogram $\boldsymbol{X}$:

$$L(\boldsymbol{X}|\boldsymbol{D}) = \prod_{f,t} p(X_{ft}|D_{ft}) = \prod_{f,t} \frac{1}{\pi D_{ft}} \exp\left(-\frac{|X_{ft}|^2}{D_{ft}}\right) \tag{5}$$

We apply the multiplicative updating rules [9] to solve the maximization. The unknown parameters Θ are initialized with random non-negative values, and 50 iterations are used in our system.

The matrix $\boldsymbol{A}_F$ resulting from the iterations is informative about the melody, since each element of it can be viewed as the intensity of a candidate f0 at a given frame. Durrieu normalizes the columns of $\boldsymbol{A}_F$, takes its elements as weights for the candidate f0's, and extracts the melody by running a Viterbi decoding on this matrix [9].

2.2. NMF-based soft masking

The iteration procedure above also gives estimates of the power spectrograms of the voice and accompaniment, which can be used to calculate the soft masks. However, after extracting the melody, we can obtain a more accurate estimate by constraining $\boldsymbol{A}_F$ to represent only the extracted melody. We run a second pass of the iteration procedure, with the elements in $\boldsymbol{A}_F$ that are far (e.g. > 0.2 semitones) from the melody line initialized with zeros. Because the updating rules are multiplicative, these zero elements will remain zeros. The soft masks are then calculated from the new estimates of the power spectrograms of the voice and accompaniment by Wiener filtering:

$$\begin{aligned} \hat{\boldsymbol{X}}_V &= \boldsymbol{D}_V./(\boldsymbol{D}_V + \boldsymbol{D}_M).*\boldsymbol{X} \\ \hat{\boldsymbol{X}}_M &= \boldsymbol{D}_M./(\boldsymbol{D}_V + \boldsymbol{D}_M).*\boldsymbol{X} \end{aligned} \quad \left(\begin{array}{c} \text{"./" denotes} \\ \text{element-wise div.} \end{array}\right) \tag{6}$$

Reversing $\hat{\boldsymbol{X}}_V$ and $\hat{\boldsymbol{X}}_M$ back into the time domain using the overlap-add method gives the separated voice and accompaniment.

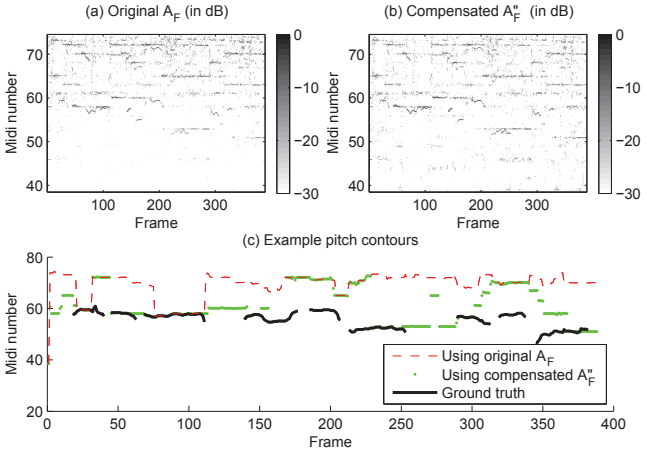

Fig. 2. The original $\boldsymbol{A}_F$, compensated $\boldsymbol{A}''_F$ and the estimated pitch contours over a 388-frame audio

2.3. Flaw of NMF-based melody extraction

As mentioned in [7], the NMF-based melody extraction method tends to estimate the pitch as one octave higher. We re-implement this algorithm and run it on the MIR-1K database [2].

First, as a basic test, we take the elements of (column-normalized) $\boldsymbol{A}_F$ directly as the weights for Viterbi decoding. We find that these octave errors happen to middle-pitched melodies, and the algorithm performs even worse for low-pitched melodies, giving wrong high-pitched estimates (see the extracted pitch contour using the original $\boldsymbol{A}_F$ in Fig. 2c). The reason for such poor performance lies in the matrix $\boldsymbol{A}_F$ produced by the first pass of iterations. One may expect that the amplitudes in each column of $\boldsymbol{A}_F$ to concentrate around the desired pitch, but we find that the amplitudes at higher frequencies in $\boldsymbol{A}_F$ are much larger than those at lower frequencies (see Fig. 2a). It is this imbalance of the values in $\boldsymbol{A}_F$ that makes Durrieu's algorithm unable to extract low-pitched and middle-pitched melodies.

We have identified two causes of the imbalance problem. First, note that the f0's of interest are chosen to be equally spaced on the midi number scale, so they are more crowded at lower frequencies on the Hertz scale. Thus, there are more power spectrum bases (i.e. columns) at lower frequencies in the $\boldsymbol{B}_F$ matrix to divide the power of each frame (see Fig. 1a), forcing the coefficients of the bases for lower frequencies to become smaller. Second, note that a power spectrum basis at a lower frequency in $\boldsymbol{B}_F$ (say, Fig. 1b) contains more harmonics and therefore more power than a basis at a higher frequency (say, Fig. 1c). This further reduces the values of the coefficients for lower frequencies in $\boldsymbol{A}_F$.

This imbalance problem underlying the NMF-based melody extraction is not well understood in [7]. An ad hoc compensation to $\boldsymbol{A}_F$ is proposed in [7] to circumvent octave errors, using modified weights for Viterbi decoding:

$$(\boldsymbol{A}'_F)_{n,t} = (\boldsymbol{A}_F)_{n,t} + 0.5(\boldsymbol{A}_F)_{n+12,t} \tag{7}$$

where n is the midi number and t is the frame number. That is, each element in $\boldsymbol{A}_F$ is compensated by adding the value at the same frame from one octave higher.

Considering the above two causes, we try to apply corresponding compensations to $\boldsymbol{A}_F$, introducing another modified weight at midi number n and frame t:

$$(\boldsymbol{A}''_F)_{n,t} = (\boldsymbol{A}_F)_{n,t} \cdot \frac{1}{f'(n)} \cdot \sum_i (\boldsymbol{B}_F)_{i,n} \tag{8}$$

Here $\frac{1}{f'(n)}$ is the reciprocal of the first derivative of Eq. (4). It represents the number of midi numbers in a unit frequency interval and describes the "crowdedness" of power spectrum bases at midi number n. The term $\sum_i (\boldsymbol{B}_F)_{i,n}$ is the total power of the power spectrum basis for midi number n.

The use of $\boldsymbol{A}''_F$ slightly alleviates the imbalance problem (Fig. 2b), but still yields a considerable number of octave errors as shown in Fig. 2c. In conclusion, the imbalance problem makes the $\boldsymbol{A}_F$ matrix not a reliable evidence for Viterbi decoding in melody extraction.

3. PROPOSED SYSTEM

The above analysis of the flaw of the NMF-based melody extraction motivates us to propose a new monaural voice and accompaniment separation system, which combines HMM-based melody extraction and NMF-based soft masking. Fig. 3 shows the flowchart of the new system. The melody extraction module is broken down in two stages: accompaniment / unvoiced / voiced (A/U/V) decision to identify the segments where the melody exists, and pitch tracking over the voiced segments. Both these stages are based on HMMs and to be introduced below. The masking stage uses Durrieu's algorithm as described in Section 2.2.

A/U/V decision – Here the acoustic features are 39-dimensional MFCC features, formed by 12 Mel-frequency cepstral coefficients (MFCC) and normalized log-energy together with their first and second differentials. Cepstral mean normalization is applied for each clip. The HMM model for A/U/V decision has three states: A, U and V. The state output distributions are modeled by 32-component diagonal-covariance Gaussian mixture models (GMMs). The HMM parameters are estimated from a labeled training database (MIR-1K).

Pitch tracking – Here the acoustic features are ESI features (Energy at Semitones of Interest). They are derived from the f0 salience function, which is a weighted sum of the magnitudes of the harmonics on the whitened spectrum [6]. Denote by $M_t(f)$ the magnitude of the whitened spectrum at frequency f at frame t, then the f0 salience function is defined as:

$$s_t(f_0) = \sum_{k=1}^{K} \frac{f_0 + \alpha}{k f_0 + \beta} M_t(k f_0) \tag{9}$$

We set the parameters as: K = 20, α = 27 Hz, β = 320 Hz. For each frame t, we calculate the salience function $s_t(n)$ for every midi number n from 38.5 to 74.5 with a step of 0.1. This results in a total of 361 salience values, which are then used to produce the final 36-dimensional ESI features by integration:

$$\mathrm{ESI}_t(n) = \sum_x s_t(x) w(x - n) \tag{10}$$

where n is a midi number from 39 to 74 with a step of 1, and $w(x - n)$ is a triangle window that extends from $n - 1$ to $n + 1$.

The HMM for pitch tracking has 36 pitch states, which correspond to the midi numbers from 39 to 74 with a step of 1. The state output distributions are modeled by 8-component diagonal-covariance GMMs. Applying the Viterbi decoding gives a coarse pitch contour that is accurate down to 1 semitone. By finding the maximizing frequency on the salience map in a 1-semitone range around the coarse pitch contour $n_1(t)$ at each frame t, we can obtain the fine pitch contour:

$$n_2(t) = \underset{n_1(t)-0.5 \le n \le n_1(t)+0.5}{\arg\max} s_t(n) \tag{11}$$

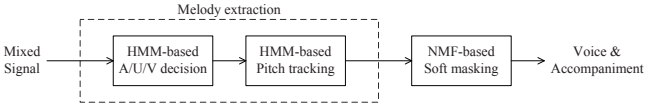

Fig. 3. Flowchart of the proposed system

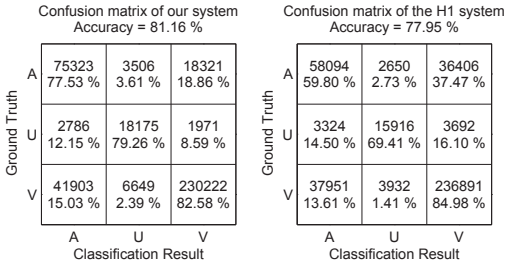

Fig. 4. A/U/V decision performance of our system and the H1 system (cited from [2]) on the MIR-1K database

4. EVALUATION AND COMPARISON [1]

4.1. Comparison with the systems of Hsu *et al.*

The H1 system of Hsu *et al.*includes an HMM-based A/U/V decision front end, but its back end uses hard masking. The A/U/V decision stage in our system is a re-implementation of the same stage in H1. The H2 system uses two streams of ESI features. Instead, we use only one stream of the newly-defined ESI feature.

The HMM training and the evaluation are carried out on the MIR-1K database with the same setup as in [2]. The database contains 1000 clips (133 minutes) of amateur singing, with A/U/V labels and f0's annotated at the frame level. The data is divided into two subsets of 487 and 513 clips for twofold cross validation.

Evaluation of A/U/V decision – Fig. 4 shows the A/U/V decision performances of our system and H1 system, in terms of confusion matrix and accuracy (the number of frames that are classified correctly divided by the total number of frames). It can be seen that the accuracy of our system is higher than that of the H1 system. Also, while the H1 system tends to classify A frames as being V, our result is more balanced. Although the two systems use the same algorithm, these performance differences may arise from the implementation details, such as the MFCC feature extraction.

Evaluation of pitch tracking – The performance of pitch tracking is measured by the overall accuracy, which is defined as the percentage of "correct" frames in all the frames. A frame is called "correct" if it is a correctly classified non-voiced (A or U) frame, or if it is a correctly classified voiced frame and the extracted pitch deviates from the true pitch by less than 1 semitone. An overall accuracy of 71.10% is reported in [4]. Our system achieves an overall accuracy of 71.57%. This indicates that our newly-defined ESI feature is better for pitch tracking than the two types of ESI features used in the H2 system.

Evaluation of overall separation performance – We use the signal-to-distortion ratio (SDR) defined in [2] as the criterion for the separation performance. The SDR, which we call "Hsu's SDR" in order to distinguish it from another SDR proposed by Durrieu, is defined as:

$$\text{Hsu's SDR} = 10 \log_{10} \frac{\langle s, \hat{s} \rangle^2}{\|s\|^2 \|\hat{s}\|^2 - \langle s, \hat{s} \rangle^2} \tag{12}$$

[1] Some separation examples are available at: `http://www.ee.tsinghua.edu.cn/~ouzhijian/maigodemo/index.htm`

Mixing ratio	H1 system			Our system	
	Ideal masks	Annot. pitch	Extr. pitch	Annot. pitch	Extr. pitch
-5 dB	10.62	7.5	-0.5	10.34	4.03
0 dB	8.36	6.0	0.9	8.70	5.31
5 dB	5.82	3.0	0.2	6.53	4.09

Table 1. Comparison of Hsu's SDR gains (in dB) on the MIR-1K database for the H1 system (cited from [2]) and our system

Clip	Original		Durrieu		Our system	
	Voice	Acc.	Voice	Acc.	Voice	Acc.
Bearlin	-5.37	5.37	6.2	11.6	3.44	8.76
Tamy	0.51	-0.51	11.5	11.0	4.17	3.66
Bent	0.01	-0.01	5.5	5.6	8.46	8.45
Chevalier	-6.79	6.79	1.5	8.3	2.72	9.50
Love	0.28	-0.28	8.6	8.4	5.17	4.89
Matter	-4.72	4.72	8.0	12.7	4.52	9.24

Table 2. Comparison of Durrieu's SDRs (in dB) for voice and accompaniment on Durrieu's database for Durrieu's system using compensated $\boldsymbol{A}'_F$ (cited from Durrieu's website) and our system

where s and $\hat{s}$ denotes the original and estimated signal respectively. The difference of the SDR before and after the separation is called the SDR gain.

We compare our system and the H1 system on the MIR-1K database at different mixing ratios of the voice and the accompaniment: -5 dB, 0 dB and 5 dB. Since the database provides annotated pitch contours, we run our system twice, using annotated and extracted pitch contours respectively. For the H1 system which uses hard masking, we also cite its SDR gains with ideal binary masks. The results are shown in Table 1. It can be seen that our system outperforms the H1 system for both cases of using annotated and extracted pitch contours. It is remarkable that the performance of our system using the annotated pitch contours comes close to or even exceeds the H1 system using ideal binary masks. This shows the advantage of NMF-based soft masking over hard masking.

4.2. Comparison with Durrieu's system

First, we compare the separation performance of our system with Durrieu's on the audio clips available on Durrieu's website[2], using the same definition of SDR as Durrieu [7]:

$$\text{Durrieu's SDR} = 10\log_{10}\frac{\|s\|^2}{\|s-\hat{s}\|^2} \tag{13}$$

The HMMs used in our system here are trained on all the 1000 clips from the MIR-1K database. The SDR results are given in Table 2. Our system performs better for some clips, while Durrieu's system is better for others. This indicates that our HMM-based melody extraction achieves comparable performance with Durrieu's NMF-based melody extraction on Durrieu's database.

Next, we run Durrieu's algorithm on the MIR-1K database using the compensated $\boldsymbol{A}''_F$ matrix. Considering that A/U/V decision is not carefully applied in Durrieu's algorithm, we compare the raw pitch accuracy (i.e. the accuracy counted on annotated voiced frames). Our system achieves a much better accuracy of 68.87% than Durrieu's 52.76%. It can also be seen from the error histograms (Fig. 5)

[2] http://perso.telecom-paristech.fr/~durrieu/en/icassp09/

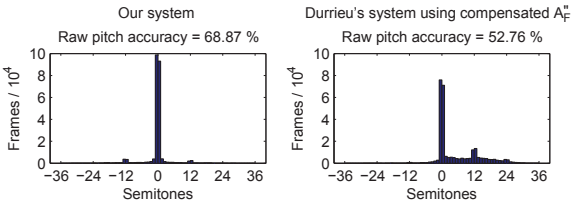

Fig. 5. Melody extraction error histograms of our system and Durrieu's system using compensated $\boldsymbol{A}''_F$

that while our melody extractor makes only a few balanced upper-octave and lower-octave errors, Durrieu's melody extractor makes a large number of various higher-pitch errors, especially upper-octave errors. This indicates that the good performance of NMF-based melody extraction obtained on the few clips on Durrieu's website cannot easily generalize to other databases, due to the imbalance problem inherent in the $\boldsymbol{A}_F$ matrix.

A final remark is that our HMM-based melody extraction achieves a significant speedup over Durrieu's NMF-based melody extraction. Our HMM-based melody extraction, which involves no iterative computation, runs $6 \sim 7$ times faster than Durrieu's.

5. CONCLUSION

In this paper, we propose a new monaural voice and accompaniment separation system, which combines HMM-based melody extraction (equipped with a newly-defined ESI feature) and NMF-based soft masking. The HMM-based melody extraction avoids the imbalance problem inherent in the NMF-based melody extraction, and runs significantly faster. Also, NMF-based soft masking gives superior performances over hard masking. Evaluations on two publicly available databases show that the proposed system reaches state-of-the-art performance and outperforms several other combinations.

6. REFERENCES

[1] Y. Li and D. L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. on ASLP*, May 2007.

[2] C. L. Hsu and J. S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Trans. on ASLP*, Feb. 2010.

[3] K. Dressler, "An auditory streaming approach on melody extraction," *Proc. MIREX*, 2006.

[4] C. L. Hsu, L. Y. Chen, J. S. R. Jang, and H. J. Li, "Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement," *Proc. ISMIR*, 2009.

[5] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative matrix spectrogram factorization in separating vocals from polyphonic music," *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Sep. 2008.

[6] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," *Proc. ISMIR*, 2006.

[7] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural music mixture de-soloing," *Proc. ICASSP*, 2009.

[8] Y. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," *Proc. ICASSP*, 2008.

[9] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," *Proc. ICASSP*, 2008.

[10] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," *Proc. European Signal Processing*, Aug. 2009.

[11] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. on ASLP*, Mar. 2010.