MODELLING AND SEPARATION OF SINGING VOICE BREATHINESS IN POLYPHONIC MIXTURES

*Ricard Marxer, Jordi Janer**

Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
ricard.marxer@upf.edu, jordi.janer@upf.edu

ABSTRACT

Most current source separation methods only target the voiced component of the singing voice. Besides the unvoiced consonant phonemes, the remaining breathiness is very noticeable to humans and it retains much of the phonetic and timbral information from the singer. We propose a low-latency method for estimating the spectrum of the breathiness component, which is taken into account when isolating the singing voice source from the mixture. The breathiness component is derived from the detected harmonic envelope in pitched vocal sounds. The separation of the voiced components is used in conjunction with an existing iterative approach based on spectrum factorization. Finally, we conduct an objective evaluation that demonstrates the separation improvement, supported also by a number of audio examples.

1. INTRODUCTION

Breathiness is an aspect of voice quality that is difficult to estimate or analyze due to its stochastic nature and wideband spectral characteristics. In western music mixture signals, this component often overlaps with other wideband components such as drums or transients. To our knowledge there are no music source separation methods that have focused on this component of the singing voice. However, in the field of speech analysis and synthesis, the decomposition and manipulation of the breathiness component has been done in a variety of areas such as text-to-speech synthesis, speech encoding, and clinical assessment of disordered voices.

For example, in [1] the authors study the relations between the vocal tract and the glottal source in human speech signals. The work in [2] focuses on the analysis of the breathy component of speech voice. It proposes a modulation-based model where the noise component of the voice is modulated by the glottal waveform. This model is used to analyse, synthesize and transform isolated voice recordings. [3] address the problem of separating the unvoiced components of the singing voice, however the authors focus on consonants and no specific breathiness models are proposed. [4] propose an extension to the source-filter model that takes into account turbulence at the glottal level and the radiation at the lips and nostrils level. The proposed model Separation of the Vocal-tract with the Liljencrants-fant model plus Noise (SVLN) shows benefits in pitch transformation and breathiness control tasks for singing voice synthesis.

All of these works focus on voice signals in isolation and do not consider the source separation problem nor the analysis of mixture signals.

* This work was supported by the Yamaha Corporation

2. PROPOSED ESTIMATION METHOD

Our method can be integrated into any source separation approach that approximates the mixture spectrum as the sum of the lead singing voice and the accompaniment spectra $\boldsymbol{V} = \boldsymbol{X}_v + \boldsymbol{X}_m$. It is appropriate for both low-latency and high-latency situations since it only requires a single audio frame.

The estimation of the breathiness component is based on the approximation of a pitched voice spectrum (with pitch f_0) as a filtered composition of two additive components: a glottal excitation $\boldsymbol{X}_v$ and a wideband component (due to the glottal air flow) $\boldsymbol{X}_{vr}$, both filtered by the vocal tract. The magnitude of the voice spectrum can be expressed in the following manner [4] (see Figure 1):

$$\begin{aligned} \boldsymbol{X}_{v'}[\omega] &= \boldsymbol{X}_v[\omega] + \boldsymbol{X}_{vr}[\omega] && (1)\\ &= \boldsymbol{L}[\omega]\boldsymbol{U}[\omega]\boldsymbol{S}[\omega]\boldsymbol{H}[\omega] + \boldsymbol{L}[\omega]\boldsymbol{U}[\omega]\gamma && (2)\\ &= \boldsymbol{L}[\omega]\boldsymbol{U}[\omega](\boldsymbol{S}[\omega]\boldsymbol{H}[\omega] + \gamma) && (3) \end{aligned}$$

where $\boldsymbol{S}[\omega]\boldsymbol{H}[\omega]$ is the spectrum of the excitation, $\boldsymbol{S}[\omega]$ is the excitation envelope, $\boldsymbol{H}[\omega]$ is a harmonic comb of unity magnitude, $\gamma\boldsymbol{U}[\omega]$ is the magnitude spectrum of the breathiness, $\boldsymbol{U}[\omega]$ is the magnitude of the frequency response of the vocal tract filter, γ is the gain of the breathiness spectrum relative to the pitched component, and $\boldsymbol{L}[\omega]$ is the component due to lips and nostrils radiation. Here we approximate the wideband component as a constant spectrum filtered by the vocal tract. This is equivalent to modeling the glottal air flow as white noise, which is realistic specially for a mid-range frequency region.

The human voice excitation envelope can be modeled, as proposed in [5], using a linear decay in the decibel/octave scale:

$$\boldsymbol{S}[\omega] = C \cdot \omega^{m/20\log_{10}(2)} \quad (4)$$

where C is a scaling factor, ω is the frequency in Hz, and m is the slope of the excitation envelope in decibels per octave (dB/octave).

In our scenario the vocal source spectrum $\boldsymbol{X}_{v'}$ is unavailable, only the mixture spectrum $\boldsymbol{V}$ is accessible. Therefore we cannot directly estimate the breathiness spectrum $\gamma\boldsymbol{U}[\omega]$ using Equation 1. Instead, we exploit the fact that at harmonic positions lf_0 of the singing voice pitch we can consider the vocals spectrum predominant $\boldsymbol{V}[lf_0] \approx \boldsymbol{X}_{v'}[lf_0]$ for all harmonic indices $l > 0$. In this work the pitch is estimated using the method presented in [6]. If we additionally consider the vocal tract filter smooth in frequency, as is done in previous works [7], we can then use interpolation between the harmonic positions to estimate the harmonic envelope $e_h[\omega] = \boldsymbol{L}[\omega]\boldsymbol{U}[\omega]\boldsymbol{S}[\omega]$ as done in [6]. By assuming the magnitudes (in the decibel scale) of $\boldsymbol{L}[\omega]\boldsymbol{U}[\omega]$ to be drawn from a

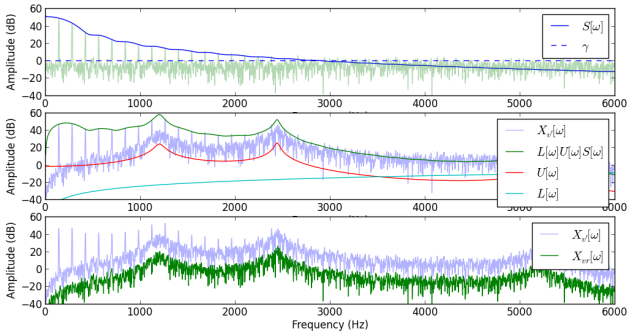

Figure 1: Representation of the different components of the singing voice model given a synthetic spectrum. Excitation spectrum containing harmonic and stochastic components (*top*). Vocals spectrum, filter model, vocal tract filter, lips and nostrils radiation filter (*center*). Vocals spectrum and breathiness spectrum (*bottom*).

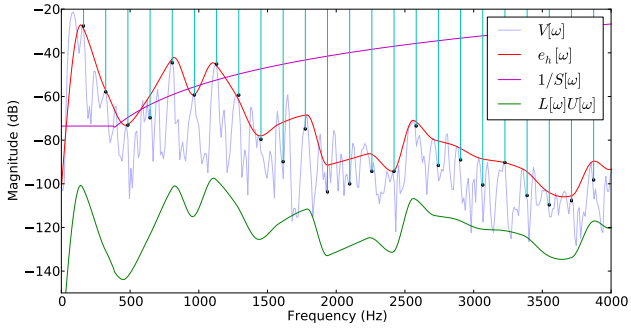

Figure 2: Breathiness estimation example. Mixture spectrum, harmonic envelope, harmonic positions, source-based whitening and the estimated breathiness.

gaussian distribution, we can make an estimation $\boldsymbol{S}[\omega]$ using least squares fitting of the model from Eq. 4 on the harmonic envelope $e_h[\omega]$. The least squares can be linear if the envelope and frequencies are first translated to logarithmic scales. This must be done on a limited region $[\omega^e_{lo}, \omega^e_{hi}]$ of the spectrum where the vocals are usually predominant and the estimated $e_h[\omega]$ reliable. Finally we whiten the harmonic envelope $e_h[\omega]$ using the excitation envelope $\boldsymbol{S}[\omega]$ derived from the excitation slope:

$$\boldsymbol{L}[\omega]\boldsymbol{U}[\omega] = e_h[\omega]/\boldsymbol{S}[\omega] \tag{5}$$

The model of Equation 4 is only valid for a mid frequency region and the estimation of $\boldsymbol{S}[\omega]$ is based on the region where the harmonics are present and predominant over the breathiness component. In order to overcome this limitation, whitening is only performed under ω^S_{hi} and is limited to $\boldsymbol{S}[\omega] = \boldsymbol{S}[\omega^S_{lo}]$ for $\omega < \omega^S_{lo}$. In our proposed method, γ is a parameter that is not estimated from the data. This parameter controls the gain of the breathiness relative to the harmonic component. In Section 4 we explore the effect of this parameter on the separation performance.

Figure 2 illustrates the intermediate results of the breathiness estimation on a spectrum of a song that contains pitched singing voice. The breathiness envelope is derived from the spectral envelope sampled at the harmonic partial frequencies.

At this point we have estimated the breathiness component as $X_{vr} = \boldsymbol{L}[\omega]\boldsymbol{U}[\omega]\gamma$. The next section describes how to use it in conjunction with an existing separation approach to obtain the final isolated singing voice.

3. INTEGRATION IN A SEPARATION APPROACH

The Smoothed Instantaneous Mixture Model (SIMM) introduced by [7] is of special interest for us, because most of the work presented here is an extension of it. We have chosen SIMM as a base due to its flexibility and simplicity. More recent methods, such as Flexible Audio Source Separation Toolbox (FASST)[8], have a more flexible and general spectrum model, however they also increase the complexity and computational cost of the process.

SIMM is an iterative parameter estimation approach, based on NMF and a source/filter model for the predominant instrument. The code implementing it is available online[1]. This method approximates the mixture spectrum as the sum of the lead singing voice and the accompaniment spectra $\hat{\boldsymbol{V}} = \boldsymbol{X}_v + \boldsymbol{X}_m$. These components are further factorized.

The accompaniment is modeled as the non-negative combination of a set of $N_{W\,m}$ constant basis components $\boldsymbol{X}_m = \boldsymbol{W}_m\boldsymbol{H}_m$. The singing voice spectrum is approximated as a multiplication of a smooth filter and a monophonic harmonic excitation $\boldsymbol{X}_v = \boldsymbol{X}_\Phi \otimes \boldsymbol{X}_{f_0}$. The factor corresponding to the filter is modeled as a combination of constant spectral shapes that are smooth in frequency $\boldsymbol{X}_\Phi = \boldsymbol{W}_\Phi\boldsymbol{H}_\Phi$. To ensure smoothness, the spectral shapes $\boldsymbol{W}_\Phi$ is modeled as a non-negative linear combination of band-limited spectra $\boldsymbol{W}_\Phi = \boldsymbol{W}_\Gamma\boldsymbol{H}_\Gamma$. The monophonicity of the excitation is achieved by modeling it as a non-negative combination of harmonic spectral templates $\boldsymbol{X}_{f_0} = \boldsymbol{W}_{f_0}\boldsymbol{H}_{f_0}$, where all the gains $\boldsymbol{H}_{f_0}$, except a region limited in frequency around the singing voice predominant pitch, are set to 0.

Some of the presented components are constant. $\boldsymbol{W}_{f_0}$ is composed of harmonic spectra with a magnitude decay computed using the Klatt glottal model. $\boldsymbol{W}_\Gamma$ is a set of band-limited filters, modeled with gaussians centered at frequencies distributed uniformly on the spectrum. In [7] a set of multiplicative update rules are derived for the other components: $\boldsymbol{H}_{f_0}$, $\boldsymbol{H}_\Phi$, $\boldsymbol{H}_m$, $\boldsymbol{H}_\Gamma$, and $\boldsymbol{W}_m$.

The reader might observe that the SIMM method already provides an estimation of the smoothed filter as $\boldsymbol{X}_\Phi$. Theoretically this filter should have a spectral shape similar to the estimated breathiness component found in section2. However, the goal of this paper was not to extend the SIMM method but to provide a general breathiness estimation valid for various separation approaches, even in low-latency conditions. The separation of the vocals is done using a Wiener filter as in [7]. The estimated breathiness spectrum is added to the estimation of the harmonic part of the voice. Using a notation similar to that used in Section 3, the estimated mixture spectrum becomes $\hat{\boldsymbol{V}} = \boldsymbol{X}_{v'} + \boldsymbol{X}_m$, which leads to the following mask:

$$\boldsymbol{m}_{v'} = \frac{\boldsymbol{X}_{v'}}{\boldsymbol{X}_{v'} + \boldsymbol{X}_m} \tag{6}$$

where $\boldsymbol{X}_{v'}[\omega] = \boldsymbol{X}_v[\omega] + \boldsymbol{L}[\omega]\boldsymbol{U}[\omega]\gamma$ is the estimated vocal source spectrum and $\boldsymbol{X}_v$ is estimated following the procedure described in Section 3. The mask is then applied to the mixture complex spectrum to compute the estimated source complex spectrum

[1] `http://durrieu.ch/phd/software.html` (last accessed on January 3, 2011)

$\widetilde{\boldsymbol{X}}_{v'} = \boldsymbol{m}_{v'} \otimes \widetilde{\boldsymbol{V}}$. Then a simple overlap-add technique is used to achieve the output waveform signal.

4. EXPERIMENTS

We prepared a dataset of multitrack recordings containing singing voice to evaluate the effect of integrating breathiness estimation into the SIMM source separation method. The multiple tracks of each recording were combined forming two sources: the vocals, and the accompaniment music created by mixing all the other tracks.

The evaluation material consists of a dataset of 14 multitrack pop-rock recordings with vocals, compiled from publicly available resources (MASS[2], SiSEC[3], BSS Oracle[4]) and 2 in-house multitrack recordings.

Measures such as SDR (Signal-to-Distortion Ratio) did not reflect the perceived differences due to the stochastic quality of the breathiness. The evaluation was done by computing the perceptually motivated measures from the well known PEASS Toolbox ([9]): OPS (Overall Perceptual Score), TPS (Target-related Perceptual Score), IPS (Interference-related Perceptual Score), APS (Artifact-related Perceptual Score).

For all the excerpts we also computed the near-optimal time-frequency mask-based separation using the BSS Oracle framework. The evaluation measures of the oracle versions of each excerpt were used as references to reduce the dependence of the results on the difficulty of each audio. Therefore the values shown are error values (lower is better) with respect to the near-optimal version.

In the experiments we set the frequency limits for the excitation slope estimation to $\omega^e_{lo} = 200$Hz and $\omega^e_{hi} = 4000$Hz. The whitening limits were set to $\omega^S_{lo} = 400$Hz and $\omega^S_{hi} = 15000$Hz. Audio examples have a sampling rate of 44.1kHz, and the spectral analysis used a frame size of 4096 without zero-padding and a hop-size of 512 samples respectively.

5. DISCUSSION

In an informal listening test we noticed that in the samples where the vocals are predominant over the background music our approach achieved its objective of maintaining the breathiness in the isolated voice. The downside, however, is that in some cases a dynamic low pass filtering is applied, which reduces the brightness of drums and cymbals in the mute version. In examples where the vocals are fast and the background is loud with relation to the vocals, the breathiness removal is less noticeable.

Looking at the objective quantitative results (not shown here), the BSSEval evaluation results show very little variation (< 0.2dB) for the different values of γ. However this does not reflect the perceived differences in the informal listening procedure. This is probably due to the fact that the differences are in frequency bands with low energy, such as the regions between the partials.

In the PEASS results (Table 1) we observed a larger change in the performance scores, however the differences in scores remained small. This could be due to limitations of the auditory model used in PEASS. Shrivastav and Sapienza[10] show the need for special care with voice breathiness quality in objective measures based on perceptual ratings.

[2] `http://www.mtg.upf.edu/static/mass`
[3] `http://sisec.wiki.irisa.fr/`
[4] `http://bass-db.gforge.inria.fr/bss_oracle/`

	APS	IPS	OPS	TPS
0.0	64.03	7.53	28.89	60.32
0.1	62.66	8.64	28.41	56.83
0.2	62.57	9.31	28.15	55.95
0.3	62.44	9.89	28.29	55.09
0.4	62.61	10.12	28.31	54.69
0.5	62.71	10.61	28.24	54.46
0.6	62.86	11.01	28.17	54.30
0.7	63.04	11.31	28.20	54.03
0.8	63.17	11.27	28.42	53.78
0.9	63.17	11.54	28.42	53.46
1.0	63.32	11.73	28.31	53.09
1.1	63.43	11.81	28.35	52.86
1.2	63.73	11.84	28.33	52.65
1.3	63.82	11.97	28.45	52.50
1.4	63.99	12.14	28.37	52.34
1.5	64.07	12.40	28.42	52.19

Table 1: Average error measures for PEASS measures for various values of γ.

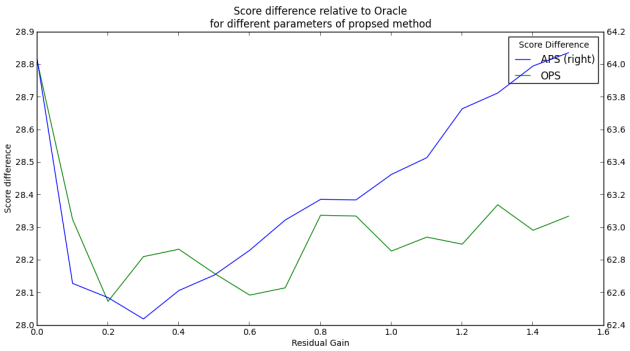

Figure 3: PEASS OPS and APS results for different parameters of the breathiness gain γ.

In any case these results reflect the conclusions extracted from the informal listening tests rather than the BSSEval results. We see a separation improvement for the OPS, APS and TPS measures, with an optimal parameter value of around $\gamma = 0.2$ for the breathiness estimation gain. While the improvement on the Overall Perceptual-related Score error is small (0.74 decrease), the proposed method does perform significantly better with respect to other measures such as APS and TPS.

Figures 3 and 4 show the trends of the different perceptually-motivated separation performance measures. Figure 4 shows the tradeoff between the interference and the target scores can be controlled with this gain parameter. OPS and APS curves in Figure 3 show a global minimum corresponding to the optimal gain for the breathiness estimation, after which the errors slowly increase with γ. The OPS curve has several local minima which could mean that the optimal value of γ depends on the song. From the results of the individual songs in Figure 5, we observe that for excerpts 5, 6 and 10 there is a clear improvement in OPS. On the other hand excerpts 2 and 12 show a significant decrease in performance when using the proposed method. Manual inspection of these in-

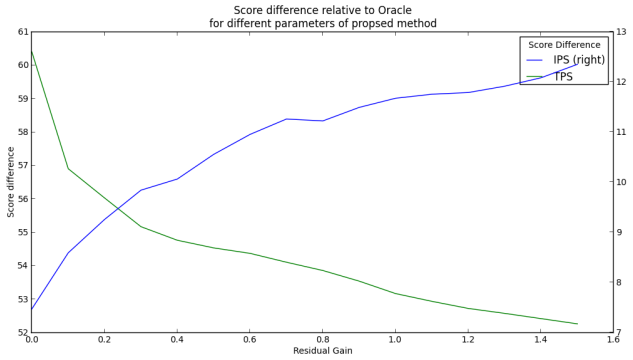

Figure 4: PEASS TPS and IPS results for different parameters of the breathiness gain.

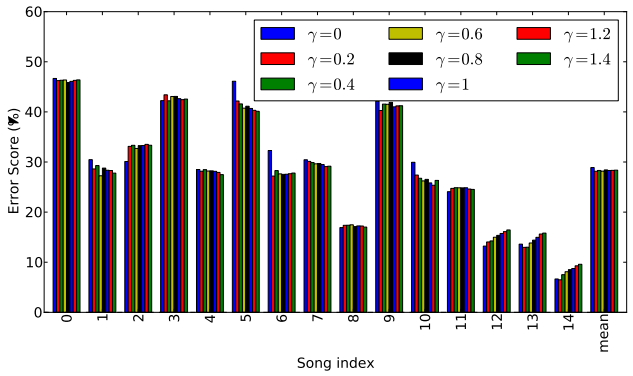

Figure 5: PEASS OPS error score (relative to Oracle) for individual songs.

stances reveal that excerpts 5 and 6 belong to the same song, with a voice containing a high degree of breathiness. Excerpt 2 shows a large number of pitch errors that could explain the large increase in errors. Finally, excerpt 12 presents a vocal track with almost no breathiness component, which would imply, as the results show, a gradual increase in errors with the increase in the parameter γ.

Another observation is that each excerpt presents a point of minimum error at a different value of γ, this shows the desirability of developing methods for estimating the optimal value γ, and thus the strength of the breathiness, from the mixture data. In a practical implementation, we suggest a user-controllable parameter γ that can be adapted to the audio content. To demonstrate the subjective improvement in the singing voice separation of our approach, we prepared a web page[5] with several audio examples.

6. CONCLUSIONS

We have proposed a method to estimate the breathiness component of the singing voice from a profesional music mixture. The method extends the source-filter model in a way similar to [4]. The spectrum model for the source is decomposed into harmonic deterministic narrowband and stochastic wideband components. The harmonic envelope of singing voice and a regression with a Klatt model is used to estimate the spectral shape of the breathiness. The breathiness is estimated up to a scaling factor, and a parameter is used to empirically control the gain of the breathiness spectrum. An experiment shows that this breathiness estimation method can be used in conjunction with the SIMM method to improve the isolation of the singing voice. Additionally, the parameter exploration of the breathiness shows that estimating the scale of the breathiness from the mixture could further improve the performance of the separation process. Future work could also be dedicated to estimating the optimal high-pass filter that models the radiation effect from lips and nostrils as well as the distribution of the glottal turbulence noise which are currently empirically parametrized.

[5] `http://www.dtic.upf.edu/~rmarxer/dafx13/breath`

7. REFERENCES

[1] K.I. Nordstrom, G. Tzanetakis, and P.F. Driessen, "Transforming Perceived Vocal Effort and Breathiness Using Adaptive Pre-Emphasis Linear Prediction," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 16, no. 6, pp. 1087–1096, 2008.

[2] D. Mehta and T.F. Quatieri, "Synthesis, analysis, and pitch modification of the breathy vowel," in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, oct. 2005, pp. 199–202.

[3] Chao-Ling Hsu and J.-S.R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 18, no. 2, pp. 310–319, 2010.

[4] G. Degottex, A. Roebel, and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE Int. Conf. on*, may 2011, pp. 5128–5131.

[5] Dennis H. Klatt and Laura C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[6] Ricard Marxer and Jordi Janer, "A Tikhonov regularization method for spectrum decomposition in low latency audio source separation," in *Proceedings IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012.

[7] Jean-Louis Durrieu, Bertrand David, and Gaël Richard, "A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation.," *J. Sel. Topics Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.

[8] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, Saint-Malo, France, Sept. 2010.

[9] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation.," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[10] Rahul Shrivastav and Christine M. Sapienza, "Objective measures of breathy voice quality obtained using an auditory model," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2217–2224, 2003.