Impact of phase estimation on single-channel speech separation based on time-frequency masking

Florian Mayer,[1,a)] Donald S. Williamson,[2] Pejman Mowlaee,[3] and DeLiang Wang[4]
[1]*FH Joanneum – University of Applied Sciences, Graz, Austria*
[2]*Department of Computer Science, Indiana University, Bloomington, Indiana 47405, USA*
[3]*Signal Processing and Speech Communication Lab, Graz University of Technology, Graz, Austria*
[4]*Department of Computer Science and Engineering, Center for Cognitive and Brain Sciences, Ohio State University, Columbus, Ohio 43210, USA*

(Received 2 June 2016; revised 7 February 2017; accepted 2 June 2017; published online 22 June 2017)

Time-frequency masking is a common solution for the single-channel source separation (SCSS) problem where the goal is to find a time-frequency mask that separates the underlying sources from an observed mixture. An estimated mask is then applied to the mixed signal to extract the desired signal. During signal reconstruction, the time-frequency–masked spectral amplitude is combined with the mixture phase. This article considers the impact of replacing the mixture spectral phase with an estimated clean spectral phase combined with the estimated magnitude spectrum using a conventional model-based approach. As the proposed phase estimator requires estimated fundamental frequency of the underlying signal from the mixture, a robust pitch estimator is proposed. The upper-bound clean phase results show the potential of phase-aware processing in single-channel source separation. Also, the experiments demonstrate that replacing the mixture phase with the estimated clean spectral phase consistently improves perceptual speech quality, predicted speech intelligibility, and source separation performance across all signal-to-noise ratio and noise scenarios. © *2017 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4986647]

[CYE] Pages: 4668–4679

I. INTRODUCTION

In many speech processing applications, the received signal is often corrupted by other interfering sources including music, background noise, or other speakers. To circumvent this problem, different source separation methods have been proposed to separate a desired source from the rest of the sources in a given observed mixture. An ideal solution in source separation is to produce separated signals with no trace of interfering sources, but also no artifacts.

Spectral phase has previously been considered unimportant for noise reduction (Wang and Lim, 1982; Vary, 1985). Because of the earlier belief on phase unimportance in audio perception, conventional single-channel source separation (SCSS) solutions are only focused on filtering the spectral amplitude of the mixture, and employ the phase of the mixed signal during signal reconstruction. This choice limits the source separation performance, as recent studies showed its positive impact and improved performance in noise reduction (Kulmer and Mowlaee, 2014; Mowlaee and Kulmer, 2015a,b; Gerkmann *et al.*, 2015; Mowlaee *et al.*, 2016a) or source separation (Gunawan and Sen, 2010; Mayer and Mowlaee, 2015; Sturmel and Daudet, 2012, 2013; LeRoux and Vincent, 2013; Magron *et al.*, 2015a). As another phase-aware source separation approach, in Bronson and Depalle (2014), complex nonnegative matrix factorization (CMF) was proposed as a tool for separating overlapping partials in mixtures of harmonic musical sources. In order to push the limited achievable performance, it is advantageous to take the phase knowledge into account in spectral amplitude modification as well as during signal reconstruction stages. In this paper, we are focused on incorporating phase information during signal reconstruction to improve the quality of the time-frequency masked (TFM) separation outcome when applied on a single-channel mixture.

Traditional speech-separation solutions are categorized into two groups (Mowlaee, 2010): (1) source-driven, represented by the computational auditory scene analysis (CASA) (Wang and Brown, 2006; Wang, 2005) and (2) model-driven (Virtanen *et al.*, 2015; Hershey *et al.*, 2010). The former group relies on processing the mixed signal directly with no prior knowledge about the underlying sources. They include a segmentation where each time-frequency unit is classified to target or masker regions followed by a grouping stage, where regions with common cues like pitch continuity and common onset/offset are assigned to the same source. The model-driven group, in contrast, relies on prior knowledge in the form of pre-trained source-specific dictionaries to capture the spectral amplitude patterns for the underlying speaker. The trained dictionaries are in the form of statistical or spectral models and can be speaker-dependent or speaker-independent depending on the training procedure applied. During separation stage, the dictionaries are searched to ascertain the best combination of the source activities and their corresponding weights. The estimated states can be used either directly to synthesize the separated signal or to

[a)]Electronic mail: florian.mayer@fh-joanneum.at.

generate a TFM (binary or soft) that is applied to the mixture to reconstruct the separated signal.

The estimated TFM is defined in the spectral amplitude domain. This mask does not modify the mixture phase, hence, it limits the achievable separation performance. The use of the mixture phase results in degraded speech quality in the separated signals (Mayer and Mowlaee, 2015; Mowlaee *et al.*, 2012a; Sturmel and Daudet, 2013; Mowlaee *et al.*, 2016a; Williamson *et al.*, 2014). Several attempts have been made recently to improve the perceived quality achievable by source separation using time-frequency masks. For example in Williamson *et al.* (2014), Williamson *et al.* present an overview study on different techniques for signal reconstruction with the focus on how to improve the speech quality of the binary masked speech. They report improved perceived quality predicted by the perceptual evaluation of speech quality (PESQ) (ITU Radiocommunication Assembly, 2001) without sacrificing the speech intelligibility predicted by short-time objective intelligibility (STOI) measure (Taal *et al.*, 2011). The same authors in Williamson *et al.* (2015) propose a two-stage procedure to improve the perceptual quality of the separated speech signal using time-frequency masking followed by estimating non-negative matrix factorization (NMF) speech model activations. They report improved perceived quality predicted by PESQ and improved intelligibility predicted by STOI. As these solutions utilize mixture phase at their signal reconstruction stage, the potential improvement due to phase processing is not addressed, hence is unknown.

Apposed to conventional SCSS methods that employ the phase of the mixed signal at the signal reconstruction stage, recently there have been several attempts to estimate the clean spectral phase of the individual sources in a mixture. Mowlaee *et al.* (2012a), proposed to use a geometric approach to first find an ambiguous set of phase candidates for the sources. The ambiguity in the pair of phase estimates is resolved by applying an additional constraint on the group delay (minimum group delay deviation) at spectral peaks found from the separated spectral amplitudes. The geometry-based approach was extended to other time-frequency constraints including across time and harmonics (Mowlaee and Saeidi, 2014). The results show improved signal reconstruction performance in PESQ and source separation criteria using the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) (Vincent *et al.*, 2006) when compared to the ideal binary mask (IBM) and multiple input spectrogram inversion (MISI) (Gunawan and Sen, 2010). Magron *et al.* (2015a) proposed phase recovery using a NMF framework. They applied phase unwrapping to ensure temporal coherence across time and frequency. The results showed that the method is suitable for a variety of pitched musical signals (Magron *et al.*, 2015a,b). Mayer and Mowlaee (2015) proposed a phase estimation method that relied on temporal smoothing of the unwrapped phase provided by a phase decomposition stage. The smoothed phase estimates were combined with the NMF and IBM-separated amplitude spectra, showing joint improvement in perceived quality and intelligibility.

The positive impact of replacing the mixture phase with an estimated clean phase has recently been reported in single-channel speech enhancement literature, showing joint improved perceived quality and speech intelligibility in speech enhancement (Gerkmann, 2014; Gerkmann and Krawczyk, 2013; Mowlaee and Kulmer, 2015a; Kulmer and Mowlaee, 2014; Mowlaee and Kulmer, 2015b). This also serves as our motivation in the current proposal for single-channel source separation. For a good recent overview of phase-aware signal processing in single-channel speech enhancement, we refer to Gerkmann *et al.* (2015) and an overview of phase-aware signal processing in speech communication we refer to Mowlaee *et al.* (2014) and Mowlaee *et al.* (2016a,b).

In this paper, we propose a phase estimation method to improve the separation performance of an estimated time-frequency masking (TFM) method that operates in the spectral amplitude domain only. The proposed method replaces the mixed spectral phase with an estimated clean spectral phase, which is used for reconstruction of the separated signals. The estimated spectral phase is calculated by temporal smoothing of the unwrapped phase, which is provided by harmonic phase decomposition of the mixture phase given the fundamental frequency of the target signal. In contrast to the preliminary work in Mayer and Mowlaee (2015), the novelty of the current work is the application of a proposed phase estimator to the estimated (not ideal) T-F masks (both binary and ratio masks) as well as a detailed analysis of the proposed pitch estimator and significance analysis of the phase-modified source separation outcomes.[1] The prior work (Mayer and Mowlaee, 2015) used phase estimation combined with ideal binary mask to demonstrate the achievable upper-bound performance.

The rest of the paper is organized as follows. Section II gives the problem definition and notations used. Section III presents an overview on the conventional TFM methods, where the mixed signal phase is used unaltered for SCSS. Section IV presents details about the proposed method to estimate the clean spectral phase for signal reconstruction. Section V presents the results, and Sec. VI concludes on the work.

II. PROBLEM DEFINITION AND NOTATIONS

Let $y(n) = x(n) + d(n)$ denote the calculation of the mixed signal in time domain from a target speech signal, $x(n)$, and the interfering noise, $d(n)$. Taking the short-time Fourier transform (STFT), we obtain $Y(k,l) = X(k,l) + D(k,l)$ where the complex STFT spectrum of the mixed signal, $Y(k, l)$, is the sum of the underlying speech and noise STFTs, with k and l referring to the frequency and time index, respectively. The STFTs are complex, containing spectral amplitude and spectral phase components, e.g., $Y(k,l) = |Y(k,l)|e^{j\phi_y(k,l)}$ with $\phi_y(k,l) = \angle Y(k,l)$, $X(k,l) = |X(k,l)|e^{j\phi_x(k,l)}$ with $\phi_x(k,l) = \angle X(k,l)$, and $D(k,l) = |D(k,l)|e^{j\phi_d(k,l)}$ with $\phi_d(k,l) = \angle D(k,l)$. The goal is to estimate the unknown desired signal, $\hat{x}(n)$, by removing the interfering source. This separated signal is reconstructed by

computing the inverse STFT of an estimated amplitude and phase,

$$\hat{x}(n) = iSTFT(|\hat{X}(k,l)|e^{j\hat{\phi}_x(k,l)}), \quad (1)$$

where $|\hat{X}(k,l)|$ is the estimated spectral amplitude of the desired signal and $\hat{\phi}_x(k,l)$ is the estimated spectral phase. Traditionally, the phase of the mixed signal is used for signal reconstruction and we have $\hat{\phi}_x(k,l) = \phi_y(k,l)$. This however, introduces artifacts of the interfering source and eventually degrades the perceived quality (Mowlaee *et al.*, 2012a). The separated spectral amplitude is estimated by applying a time-frequency mask denoted by $\hat{M}(k,l)$ to the mixture spectral amplitude $|Y(k,l)|$ given by

$$|\hat{X}(k,l)| = \hat{M}(k,l)|Y(k,l)|, \quad (2)$$

where $\hat{M}(k,l)$ can be a binary mask [i.e., $\hat{M}(k,l) = \widehat{\text{IBM}}(k,l)$] or a ratio mask [i.e., $\hat{M}(k,l) = \widehat{\text{IRM}}(k,l)$]. Further details on how to estimate the IBM and IRM are given in Sec. III.

III. CONVENTIONAL TIME-FREQUENCY MASKING

Time-frequency masking is an effective approach for source separation, where two types of time-frequency masks are commonly used: binary and ratio. The IBM is a two-dimensional binary matrix used to label T-F units of a mixture as noise or speech dominant (Wang, 2005), where T-F units with the value of 1 are deemed speech dominant and retained, and noise dominant T-F units are removed. Given the T-F representation of the speech, $X(k, l)$, and noise, $D(k, l)$, the IBM is defined as follows:

$$\text{IBM}(k,l) = \begin{cases} 1, & \text{if } |X(k,l)| > |D(k,l)| \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

An estimate of the speech signal is generated by applying the estimated IBM to the T-F representation of the mixture. Recent studies, that use binary masks to address monaural speech separation in noisy environments, have shown improved speech intelligibility (Kim *et al.*, 2009; Healy *et al.*, 2013), but the resulting speech quality is still not satisfactory (Williamson *et al.*, 2015). Applying the estimated IBM for SCSS, results in errors due to the removal of certain portions of the speech and retaining portions of the noise miss-classified as speech, hence, degrading the perceived quality.

An alternative to binary masking is a continuous (ratio) mask. Before describing this mask, let us consider the geometry of source separation in the power-spectrum domain,

$$|Y(k,l)|^2 = |X(k,l)|^2 + |D(k,l)|^2 + 2|X(k,l)||D(k,l)|\cos(\theta), \quad (4)$$

where $\theta = \phi_x(k,l) - \phi_d(k,l)$ is the phase difference between the two sources. We assume that the phase difference is uniformly distributed and the two sources are uncorrelated. This brings us directly to the assumption, that the sum of the power spectra of two uncorrelated sources result in the power spectrum of the mixture signal. Further approximation of Eq. (4) yields a mask, in contrast to NMF or related soft-masks, but similar to the square root Wiener filter defined in Loizou (2013),

$$H(k,l) = \sqrt{\frac{P_{xx}(k,l)}{P_{xx}(k,l) + P_{dd}(k,l)}}, \quad (5)$$

where P_{xx} and P_{dd} denote power spectral densities for speech and noise, respectively, and are approximated by their short-term magnitude spectrum squared. Similarly in Févotte and Godsill (2005), the authors proposed an IRM for audio source separation, taking into account the spectral magnitude ratio. The IRM is a two-dimensional matrix where each T-F unit represents the proportion of energy that is attributed to speech. Therefore, IRM-separated speech retains the correct amount of speech energy in each T-F unit. The IRM has values in the range of [0,1] and is defined in Wang *et al.* (2014) as follows:

$$\text{IRM}(k,l) = \left(\frac{|X(k,l)|^2}{|X(k,l)|^2 + |D(k,l)|^2}\right)^{1/2}. \quad (6)$$

The IRM has been shown to be more effective for DNN based estimation than other training targets such as the IBM, cochleagram, and spectrogram in terms of achievable improvement in the perceived quality and intelligibility (Wang *et al.*, 2014).

In the current work, separate DNNs are used to estimate the IBM and IRM for a given mixture. The DNNs are trained from the set of complementary features that is extracted from a 64-channel gammatone filterbank: amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), and Mel-frequency cepstral coefficients (MFCC), as well as their deltas (Wang *et al.*, 2013). The DNNs are identical in terms of structure and parameter values. A sliding context window is used to splice adjacent frames into a single vector for each time frame. This is employed for the input features and target labels of the DNNs. In other words, the DNN maps a set of frames of features to a set of frames of labels for each time frame (Zhang and Wang, 2014; Wang *et al.*, 2014). Useful information is carried across time frames, and a context window allows the DNN to incorporate this information. During testing, the set of estimated labels for each time frame is unwrapped and averaged in order to obtain estimates of the IBM and IRM. As illustrated in Fig. 1(a), the estimated masks are applied to the cochleagram of the mixed signal to produce a speech estimate.

IV. PROPOSED PHASE ESTIMATOR

In this section, we present details about the proposed phase estimation approach. The method relies on harmonic phase decomposition, first proposed in Degottex and Erro (2014), which itself demands access to the pitch trajectories of the underlying sources in the mixed signal. The obtained

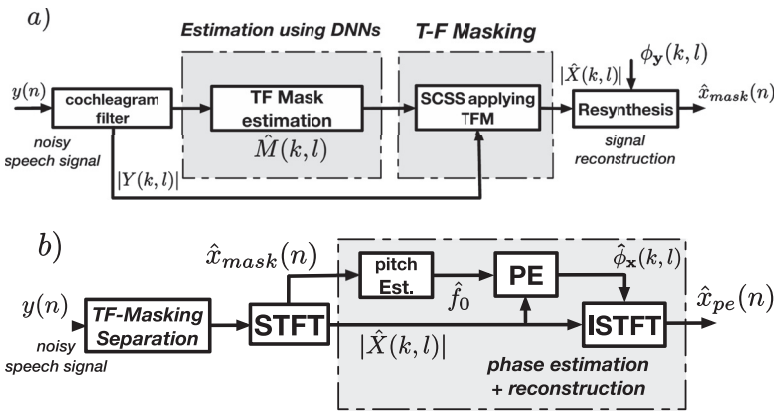

FIG. 1. (a) *Conventional time-frequency masking (TFM)-SCSS methods* (Ideal Binary Mask and Ideal Ratio Mask), and (b) proposed method to incorporate phase information at signal reconstruction in TFM-SCSS leading to $\hat{x}_{\mathrm{pe}}(n)$.

noise robust f_0-estimates are then used to extract an unwrapped harmonic phase, followed by temporal smoothing to provide an enhanced spectral phase for reconstructing the separated signals.

A. Phase estimation

Figure 1(b) illustrates the principle of the proposed phase-aware TFM method. After TF-masking, the f_0 trajectories of the separated speech signals are required to be used as prior information for the subsequent phase estimation step. Given the mixed signal, our goal here is to estimate the clean spectral phase denoted by $\hat{\phi}_x(k,l)$ using the estimated fundamental frequency. The enhanced phase obtained by temporal smoothing of unwrapped phase is then used for synthesis, where we replace the phase obtained from the mixture, with the estimated one. The steps required for phase estimation are explained in detail in the following.

1. Phase decomposition

The spectral phase at harmonics alters at multiples of 2π due to phase wrapping. To circumvent these discontinuities, we apply the phase decomposition from Degottex and Erro (2014) to extract an unwrapped harmonic phase, known for its smooth changes across time within speech regions. In order to reduce the error variance, the linear phase, which wraps the harmonic phase across time according to its harmonics, needs to be removed from the harmonic phase. Applying a temporal smoothing filter on the resulting unwrapped phase enables us to reduce the impact of noise. As the linear phase is caused by cyclic wrapping of the fundamental frequency, having access to a reliable pitch estimate plays a key role in the ultimate accuracy of the phase estimation procedure described here. The pitch estimation will be discussed in Sec. IV A 2.

Let $f_0(l)$ represent the fundamental frequency of the desired source at the lth frame, provided by a pitch estimator (see Sec. IV A 2 for further details). In the harmonic model, phase distortion (Degottex and Erro, 2014) is applied to access different harmonic components of the phase described in the following. Using $f_0(l)$, pitch-synchronous time-segmentation is applied to the time-domain TFM separated signal denoted by $\hat{x}_{mask}(n)$. The segmentation is characterized by time stamps $t(l) = t(l-1) + 1/[4f_0(l-1)]$, where $t(l)$ is defined as the time instant at the lth frame. This results in time-segments $\hat{x}_h(n',l) = \hat{x}_{mask}(n' + t(l))w(n')$ where $w(n')$ is the analysis window of the length N and $n' \in [-(N-1)/2, (N-1)/2]$. For phase estimation, a careful consideration about the window type and length is necessary. As reported in Mowlaee and Kulmer (2015b),

$$\hat{x}_h(n',l) \approx \sum_{h=1}^{H} a(h,l)\cos(h\omega_0(l)n' + \psi(h,l))w(n'), \quad (7)$$

where $a(h,\ l)$ and $\psi(h,l)$ are the harmonic amplitude and phase, respectively, h is the harmonic index, and H is the harmonic model order. In voiced regions, each time-segment $\hat{x}_h(n',l)$ can be approximated as a sum of harmonics consisting of amplitude $a(h,\ l)$ and harmonic phase $\psi(h,l)$, shown in Eq. (7). This assumption is due to the speech production model assumed in the harmonic model plus phase distortion (HMPD), proposed in Degottex and Erro (2014). We extract the phase information $\psi(h,l)$ defined as the phase at the hth harmonic and lth frame index. As shown in Mowlaee and Kulmer (2015a) the harmonic phase can be decomposed into three components:

$$\psi(h,l) = \underbrace{\angle V(h,l) + \psi_d(h,l)}_{\text{Unwrapped phase: } \Psi(h,l)} + \underbrace{h\sum_{l'=0}^{l}\omega_0(l')(t(l') - t(l'-1))}_{\text{Linear phase: } \psi_{\mathrm{lin}}(h,l)}. \quad (8)$$

The *minimum phase spectrum* represents the first term of Eq. (8), and it is attributed to the vocal tract filter. The *dispersion phase*, or source shape, $\psi_d(h,l)$ denotes the second term and characterizes the pulse shape. It also captures the stochastic characteristic of the phase at harmonic h and, as shown in Degottex and Erro (2014), is known to smoothly evolve across time. The *linear phase*, the last term, wraps the harmonic phase across time according to h, normalized fundamental frequency $\omega_0(l)$ and the time interval between two consequent frames $t(l) - t(l-1)$. When the signal is corrupted with an interfering signal or noise, the unwrapped phase component becomes corrupted as well.

We define $\Psi(h,l)$, the unwrapped phase, as the combination of the minimum phase spectrum and dispersion phase,

therefore the harmonic phase is given by the superposition of the unwrapped phase and the linear phase

$$\psi(h,l) = \Psi(h,l) + \psi_{\text{lin}}(h,l). \tag{9}$$

In our proposed phase estimation method, given the fundamental frequency estimate we propose to approximate the linear phase contribution between consequent frames denoted by $\hat{\psi}_{\text{lin}}(h,l)$. We further define $\hat{\omega}_0(l) = 2\pi\hat{f}_0(l)/f_s$ with $\hat{f}_0(l)$ as the estimated fundamental frequency for the lth frame and f_s as the sampling frequency. Then the linear phase estimate is given by

$$\hat{\psi}_{\text{lin}}(h,l) = \sum_{l'}^{l} h\hat{\omega}_0(l')(t(l') - t(l'-1)). \tag{10}$$

This estimate is used to find the unwrapped harmonic phase which is given by subtracting the linear phase from the harmonic phase, thus we have

$$\hat{\Psi}(h,l) = \psi(h,l) - \hat{\psi}_{\text{lin}}(h,l). \tag{11}$$

2. Robust pitch estimation using frequency histograms (RPEFH)

To perform accurate phase estimation, a smooth trajectory of the fundamental frequency f_0, with the least number of f_0 outliers is desired. The outliers in the f_0 estimate are introduced by the interfering signal in the mixture. Fundamental frequency estimation becomes difficult when the underlying frequency trajectories of the target and interfering signal are close together, hence, the fundamental frequency of the target signal is hard to distinguish. To address the multi-pitch estimation problem, we propose to incorporate the histogram of the fundamental frequencies, obtained from the TFM-separated signals (using estimated IBMs and IRMs). Figure 2 shows the block diagram of the pitch estimator used in this work.

The initial pitch estimate is provided using the pitch estimation filter with amplitude compression (PEFAC) proposed in Gonzales and Brookes (2014). To estimate the frequency trajectory of the desired source, hypotheses of the minimum and maximum occurring f_0 value (f_{min} and f_{max}) have to be provided as additional inputs to the PEFAC algorithm. Initially we assume a frequency range of 50 to 500 Hz, which covers the possible pitch range for speech. The histogram analysis reveals that the estimated f_0 trajectory from the mixed signal, $\hat{f}_0$, consists of values that are not within the confidence interval ($\mu \pm 1.96\sigma$) of the desired pitch information. These outliers, caused by the interfering signal

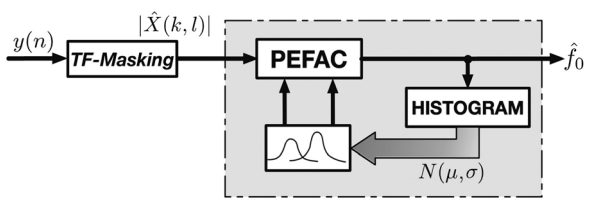

FIG. 2. Block diagram of the fundamental frequency estimator.

in the mixture, cause errors in the linear phase approximation required in Eq. (8). Statistical analysis is performed on the estimated trajectories to characterize the mean, μ, and the variance, σ_{f_0}, of the underlying pitch distribution. To this end, we fit a Gaussian probability density function to $\hat{f}_0$, where its mean, μ, is assumed to be within the valid pitch range of $[f_{0_{\min}}, f_{0_{\max}}]$ and we have

$$\hat{f}_0 \sim N(\mu, \sigma_{f_0}) \quad \text{where } \mu \in [f_{0_{\min}}, f_{0_{\max}}]. \tag{12}$$

As a result we ascertain an updated pitch range of $f_{0_{\min}}$ and $f_{0_{\max}}$ as well as a new voicing probability from PEFAC. Incorporating this statistical framework for fundamental frequency estimation allows to further incorporate the renewed signal characteristics, to re-apply the pitch estimation step on the TFM-separated signal $\hat{x}(n)$ for a second time. As the TFM-separated signal inherits less trace of the interfering masking signal, an improved f_0 estimate, hence, a more reliable linear phase is expected at this stage. In Sec. IV A 3, we demonstrate the effectiveness of the proposed pitch estimator for single-channel source separation and phase enhancement for signal reconstruction.

3. Temporal smoothing of unwrapped phase

A low variance of the spectral phase at harmonics is a desired characteristic for high quality synthesized speech (Koutsogiannaki *et al.*, 2014). Further, recent advances in phase-aware processing for speech enhancement (Mowlaee and Kulmer, 2015b; Kulmer and Mowlaee, 2014; Mowlaee and Kulmer, 2015a) report improved signal reconstruction and noise reduction when the variance of estimated phase at harmonics is reduced. Therefore, in the current contribution, we apply temporal smoothing filters on the unwrapped phase, in order to reduce the large circular variance caused by interferences but also artifacts.

From circular statistics, we compute the circular mean of $\hat{\Psi}(h,l)$ by averaging out the contribution of the interfering signal applied on voiced frames-only. A smoothed unwrapped phase estimate is given by

$$\bar{\Psi}(h,l) = \angle \frac{1}{|\mathcal{R}|} \sum_{l' \in \mathcal{R}} e^{j\hat{\Psi}(h,l')}, \tag{13}$$

where $\mathcal{R}$ is the set of neighboring frames lying within the time span of 20 milliseconds. This balances the trade-off between getting reliable statistical estimates and fulfilling the stationarity of processing speech within short enough frames.

B. Improved signal reconstruction using enhanced spectral phase

By adding back the approximated linear phase, provided in Eq. (10), to the temporally smoothed unwrapped harmonic phase, $\bar{\Psi}(h,l)$, an enhanced harmonic phase is provided,

$$\hat{\psi}(h,l) = \hat{\psi}_{\text{lin}}(h,l) + \bar{\Psi}(h,l). \tag{14}$$

To combine the estimated phase with the spectral amplitude, obtained by applying an estimated IBM or IRM computed in the STFT-domain, we need to transform the estimated

enhanced phase from harmonic domain back to the STFT domain. To this end, Magron *et al.* (2015b) introduced *regions of influence* to ensure that the phase at a given frequency channel is unwrapped appropriately to its instantaneous frequency. They estimated an instantaneous attack time n_0, by taking into account the instantaneous frequency used for unwrapping over time (temporal). This leads to the estimation of n_0 in each frequency channel in order to perform an unwrapping along frequency bins (spectral).

In this work, the STFT-frequency bins k within the mainlobe width N_p, located adjacent to the harmonic multiples $h\hat{f}_0$, are replaced by $\hat{\psi}(h,l)$. We assume that the underlying harmonics are well separated in frequency domain given a long enough frame length. Further, we transform the estimated harmonic phase into STFT domain by modifying the enhanced spectral phase at frame l and frequency bin k within the mainlobe width N_p of the analysis window

$$\hat{\phi}_x(\lfloor h\omega_0 N_{\mathrm{DFT}}/(2\pi)\rfloor + i, l) = \hat{\psi}(h,l),\ \forall i \in [-N_p/2, N_p/2],$$
$$N_p = \Delta\omega_{MW} N_{\mathrm{DFT}}/(2\pi) \tag{15}$$

with N_{DFT} as the DFT size and $\Delta\omega_{MW}$ as the mainlobe bandwidth. The estimated STFT phase is combined with the estimated source spectral amplitude denoted by $|\hat{X}(k,l)|$. The lth segment of the phase-enhanced separated signal is then given by

$$\hat{x}_l(n) = \mathrm{DFT}^{-1}\left\{|\hat{X}(k,l)|e^{j\hat{\phi}_x(k,l)}\right\}. \tag{16}$$

Applying overlap-add on $\hat{x}_l(n)$, for all l, reveals the phase-enhanced time-frequency masked signal $\hat{x}_{\mathrm{pe}}(n)$.

V. EXPERIMENTS

We demonstrate the effectiveness of the proposed phase estimation method for single-channel source separation. After a description regarding the chosen experiment setup, we provide a detailed analysis of the pitch estimation used as input for phase estimator. A proof-of-concept demonstrates the impact of phase modification on the source separation outcome. Finally, we report the averaged results (on blind scenario) with comparison to benchmark methods.

A. Experiment setup

1. Experiment one: Speech in noise

We use 550 utterances from the IEEE corpus (IEEE Audio and Electroacoustics Group, 1969) as our training utterances. The IEEE corpus consists of 720 utterances read by a single male speaker. The testing set consists of 60 randomly chosen utterances from the IEEE corpus which are disjoint from the utterances used for training. Each signal is down-sampled to 16 kHz. The training utterances are mixed with speech-shaped noise (SSN), cafeteria, factory, and babble noise. SSN is stationary, while the other selected noise types are non-stationary. The noisy test files are produced by mixing the clean speech utterances with the aforementioned noise files at signal-to-noise ratios (SNRs) of −6, −3, 0, 3, and 6 decibels.

Each noise is around 4 min long. To create the training sets, we use random cuts from the first 2 min of each noise to mix with the training utterances. The test mixtures are constructed by mixing random cuts from the last 2 min of each noise with the test utterances. Dividing the noises into two halves ensures that the testing noise segments are unseen during training.

The DNNs used for estimating the IBM and IRM have three hidden layers, each having 1024 rectified linear hidden units (ReLU) (Glorot *et al.*, 2011). The standard backpropagation algorithm coupled with the mean-square error cost function is used to train the DNNs. A momentum term and the adaptive gradient descent (Duchi *et al.*, 2011) are used for optimization. A momentum rate of 0.5 is used for the first 5 epochs, after which the rate changes to 0.9. A sigmoid activation function is used in the output layer. A context window that spans five frames centered at the current frame is used to splice the input features, while the length of the context window is three frames for the DNN training targets.

The separation results are reported in terms of blind source separation evaluation (BSS EVAL) measures (Vincent *et al.*, 2006) standardized for speech quality estimation of source separation algorithms. BSS EVAL toolkit is composed of three SNR measures: signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) all measured in decibels. STOI (Taal *et al.*, 2011) and PESQ (ITU Radiocommunication Assembly, 2001) are also used to predict the speech intelligibility and perceived quality, respectively, where both show a high correlation with subjective tests in single-channel source separation (Mowlaee *et al.*, 2012b).

Furthermore for the phase estimation setup, the masked-speech signals are downsampled to 8 kHz. A Blackman window [minimizing spectral leakage (Harris, 1978) also reported in Mowlaee and Kulmer (2015b)] with length of 32 ms is used. For the temporal smoothing step in the proposed phase estimation method [see Eq. (13)], a moving average filter with a filter length R of 20 milliseconds is used. This decision is based on finding a reasonable trade off between a low circular phase variance and preserving the short-time speech stationarity.

2. Experiment two: Speech mixture in noise

This experiment evaluates separation performance when two speakers and noise are present in a mixture. The two speaker mixtures are generated by combining IEEE utterances from a single male and a single female speaker with SSN, cafeteria, babble, and factory noise. The speech-to-speech ratio (SSR) between the male and female utterances is set to −3, 0, or 3 dB. The two-speech mixed signal is combined with random cuts of noise at a 10 dB SNR. Four separate DNNs are used to extract the male or female speech by estimating the IBM or IRM. A total of 14 700 mixtures (35 male utterances, 35 female utterances, 3 SSRs, and 4 noises) are used to train each DNN. The features and network configuration are as described in experiment 1. Each DNN is evaluated with a testing set that consists of 1200 mixtures (10 male utterances, 10 female utterances, 3 SSRs,

and 4 noises), where the utterances and random cuts of noise are different than those used during training. An NMF algorithm, proposed in Virtanen *et al.* (2013), was used as a benchmark method. To create the target and masker NMF dictionaries we chose 35 sentences for each speaker for training speaker dependent dictionaries.

B. Performance evaluation of the proposed f_0 estimator

To assess the effectiveness of the proposed histogram-based pitch estimator, described in Sec. IV A 2, we quantify its performance in terms of the Gross Pitch Error (GPE) and the Fine Pitch Error (FPE) proposed in Chu and Alwan (2009) and Babacan *et al.* (2013). GPE represents the frames considered voiced by the estimator and the ground truth, for which the error is higher than 20%. The FPE is the standard deviation of the error from frames without Gross Pitch Error. In this experiment we used 15 sentences from IEEE corpus, mixed with noise files at different SNRs.

Table I shows the evaluation results for GPE and FPE. As upper-bound scenario, denoted as "UB," we report the estimated fundamental frequency from the TFM-separated signal combined with clean spectral phase. As lower-bound, denoted as "LB," we report the results for the f_0 estimates from the observed mixed signal. The estimates for both scenarios are obtained using PEFAC, addressing the question regarding the importance of the clean spectral phase for a reduced fundamental frequency estimation error. As a benchmark, we compare the PEFAC f_0 estimate of the estimated TFM signals denoted as "est. TFM (mixture phase)" with the f_0 estimate obtained from the same signal using the proposed pitch estimator denoted as "est. TFM (proposed)." For all SNRs, the TFM separated signals provide an improved GPE and FPE compared to the mixture. The results also show an improved GPE obtained by the clean spectral phase or the proposed f_0 estimation compared to TFM and scenarios which makes use of the mixed signal phase. In terms of FPE, for −3 and 6 dB SNRs, the IBM with mixed signal phase provides less error than the proposed method. For IRM scenario, the proposed pitch estimator outperforms the mixed signal phase scenario for all SNRs. This suggests the advantages of incorporating a soft mask rather than a binary mask for pitch estimation. The pitch evaluation results reveal, that the use of a proper frequency range, obtained by the statistical analysis, significantly improves the pitch estimation accuracy.

C. Proof-of-concept experiment

We consider two proof-of-concept experiments to demonstrate the usefulness of the proposed phase estimator when combined with the estimated time-frequency mask signals. The results are shown in Fig. 3(a) and 3(b) for estimated IBM and IRM scenarios, respectively.

As our first experiment, we consider a male speaker saying "*The small red neon lamp went out*" mixed with SSN noise file at 0 dB SNR. The results are shown in Fig. 3 for the estimated IBM and different phase outcomes in terms of (top) spectrogram, (middle) group delay, and (bottom) phase variance. From left to right, the outcomes for clean reference, mixed signal, time-frequency masked (TFM) signal using mixture phase, and proposed phase-enhanced TFM signal are shown. From the spectrograms of the mixed signal, shown on the first row, SSN widely masks the desired speech signal, in particular at fundamental frequency and lower harmonics. The estimated IBM is capable of reconstructing certain time-frequency regions of the target signal.

The proposed method contributes to further emphasize the desired harmonic structure in the separated signal. The group delay plot (second row) reveals that the binary TFM, when combined with enhanced spectral phase, is capable of recovering certain harmonic structure across frequencies. The circular variance (bottom row) shows a reduction of the large variance of the mixed signal in the phase-enhanced reconstructed IBM.

The results from the estimated IRM are shown in Fig. 3(b). The mixed signal is produced by mixing a male target speaker saying "*The fan whirled its round blades softly*" with cafeteria noise at 0 dB SNR. From the spectrograms (first row), the estimated IRM, compared to IBM, provides smoother transitions between the harmonics. Some masker components are removed via applying the proposed phase estimation method. Comparing the patterns to clean and mixture phase outcomes, group delay and circular phase variance shows further improvement after applying the proposed phase estimator.[2]

TABLE I. Evaluation of the proposed fundamental frequency estimator compared to the clean phase upper-bound (UB), the lower bound mixture (LB), as well as the f_0 estimate from the input signal (mixture phase) and the proposed f_0 estimator outcome (phase-enhanced).

	Gross-Pitch Error (GPE) (%)					Fine-Pitch Error (FPE) (Hz)				
	SNR (dB)					SNR (dB)				
GPE/FPE input signal	−6	−3	0	3	6	−6	−3	0	3	6
(UB): est. BM (PEFAC)	26.07	23.67	21.10	19.41	18.56	0.71	0.59	0.75	0.63	0.69
(UB): est. RM (PEFAC)	25.16	22.29	18.51	18.62	16.54	0.71	0.63	0.79	0.74	0.88
est. BM (PEFAC)	48.01	39.75	32.13	28.22	23.26	1.49	0.84	0.86	0.92	0.69
est. BM (proposed PE)	39.37	33.77	27.96	25.05	21.90	1.25	0.85	0.80	0.87	0.88
est. RM (PEFAC)	52.45	44.44	37.80	31.98	27.65	1.99	1.09	1.26	0.89	0.85
est. RM (proposed PE)	46.32	39.21	32.13	28.89	25.34	1.46	0.97	0.94	0.83	0.69
(LB): Mixed signal (PEFAC)	66.2	60.55	53.98	46.52	40.33	2.96	2.42	2.22	1.71	1.46

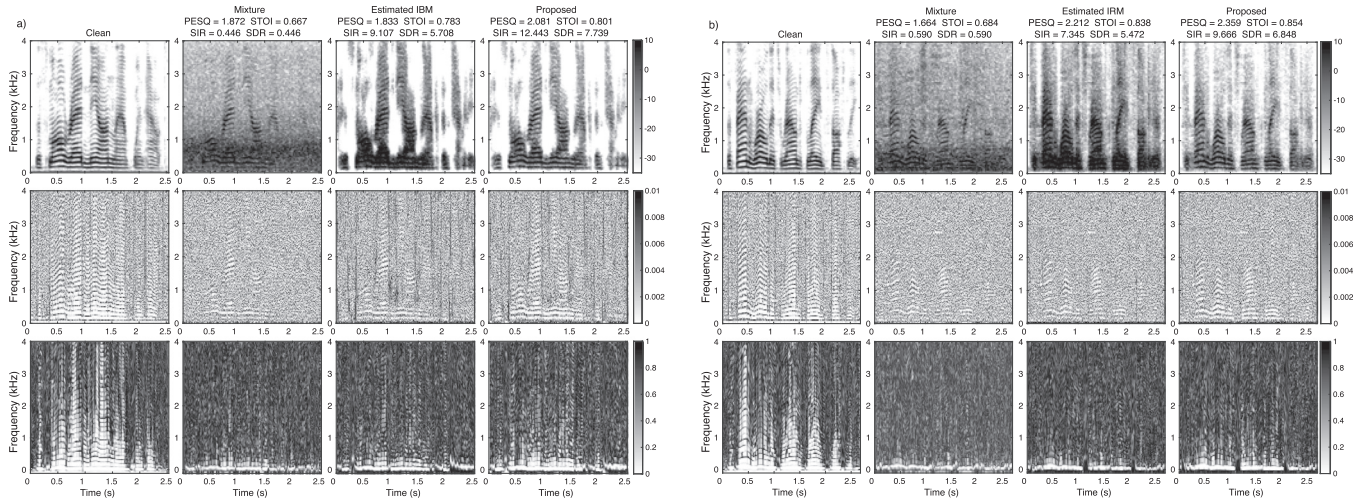

FIG. 3. Proof-of-concept experiment for estimated (a) IBM and (b) IRM method mixed at 0 dB shown as (top) spectrogram, (middle) group delay, and (bottom) phase variance. The results for clean (first), mixture (second), estimated IBM/IRM (third) are shown for comparison.

D. Source separation results

1. Experiment one: Speech in noise

Here, we report the results for the first experiment averaged on the whole utterances and noise types at different SNRs. Figure 4 and Fig. 5 show the bar plots to quantify the effectiveness of the proposed phase-enhanced TFM compared to conventional TFM using the phase of the mixed signal. As upper-bound, we include the outcome when clean spectral phase is used, which delivers the achievable upper-bound performance by a phase estimation procedure combined with the spectral amplitude provided by TFM separated signal. As a benchmark, we added two phase enhancement methods that rely on Griffin and Lim iterative signal reconstruction procedure (Griffin and Lim, 1984): partial phase reconstruction (PPR) (Sturmel and Daudet, 2012) and informed source separation using iterative reconstruction (ISSIR) (Sturmel and Daudet, 2013). The two methods incorporate the TFMs as confidence domains to guide the iterative signal reconstruction in an iterative way.

The following observations are made:

- In terms of objective speech quality and intelligibility, predicted by PESQ and STOI, a consistent improvement for all SNRs in the range of −6 dB to 6 dB is achieved when the estimated phase is used for IBM and IRM rather than the mixed signal phase. The proposed method also leads to a reasonable advantage in SIR and SDR, which is obtained even for low and high SNRs. In this experiment, the SAR results are not improved by employing the estimated phase. This could be due to the fact that the processing of the noisy signal introduces new artifacts. The SAR scores due to phase modification are still close to those obtained for noisy phase TF enhanced signal.
- The clean spectral phase results indicate the clear positive impact of incorporating an enhanced spectral phase when combined with the TFM separated spectral amplitude.
- The white bars show the ideal binary and ratio mask results, known as the upper-bounds for the achievable separation performance by a TFM. It is a crucial finding that

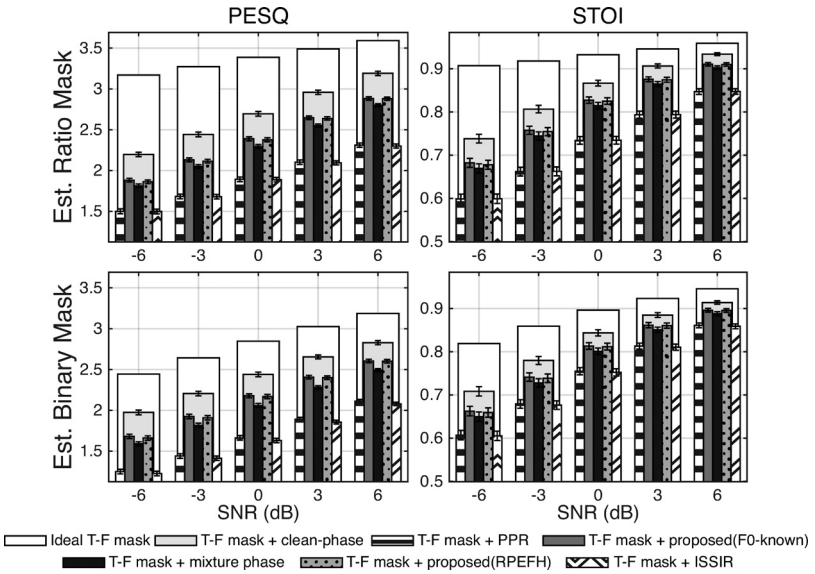

FIG. 4. Evaluation results: Estimated (top) IRM and (bottom) IBM in terms of quality (PESQ) and intelligibility (STOI). Results for oracle phase and IBM/IRM (mixture phase) are reported for comparison.

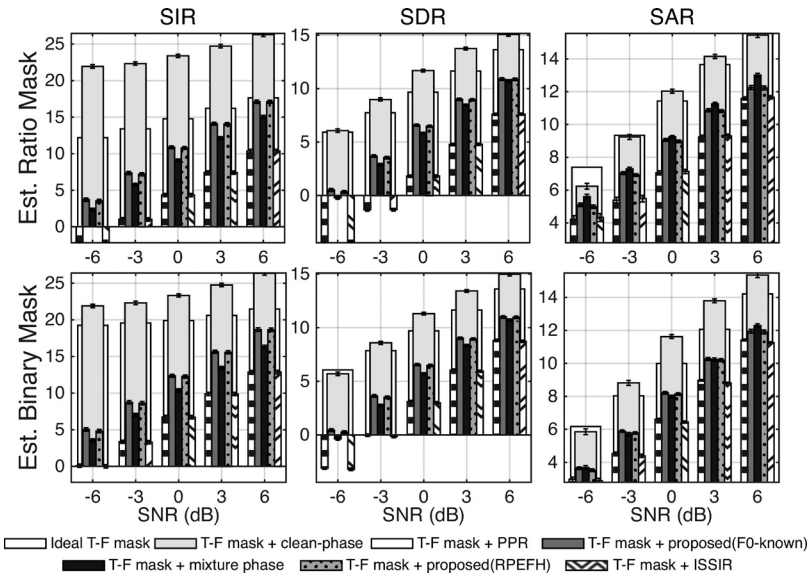

FIG. 5. Evaluation results: Estimated (top) IRM and (bottom) IBM in terms of separation performance (SIR), (SDR), and (SAR). Results for oracle phase and IBM/IRM (mixture phase) are reported for comparison.

the gap between the clean-phase upper-bound and the ideal TFM-separated signal is still considerable for PESQ and STOI, this might be due to the large gap between the ideal and estimated TFM-separated signal.

- PPR and ISSIR results achieve relatively poor separation performance which could be due to their sensitivity to errors introduced by the estimated spectral amplitudes using estimated IBM or IRM. These errors are re-introduced within each iteration when combining the spectral amplitudes with newly estimated spectral phase. These observations demonstrate that the proposed method is less prone to errors of the estimated spectral amplitude, in contrast to the GL-based iterative phase recovery methods (Sturmel and Daudet, 2012, 2013).
- The remaining gap between the estimated phase and the clean phase emphasizes on the importance of having a reliable phase estimate in SCSS. This observation also explains the importance of phase in SCSS, encouraging for further studies pushing the limits in conventional single-channel source separation methods relying on mixed spectral amplitude modification only.
- To show that the proposed method is robust to fundamental frequency estimation errors, we also include the results obtained for the f_0-known scenario, where the f_0 trajectories are ascertained from the clean signal using PEFAC. The comparison of f_0-known and proposed f_0-estimated [referred to as Robust Pitch Estimation using Frequency Histograms (RPEFH)] scenarios reveals that the proposed phase estimation method is not that sensitive to the errors introduced by the pitch estimator, justified by the similar performance in both f_0 scenarios.

Finally, we conducted two-paired Kolmogorov-Smirnov (KS) tests to assess the significance of the reported results due to the spectral phase modification in TFM compared to conventional TFM using mixed signal phase. For this test, in order to provide a comprehensive analysis independent of the applied SNR, each evaluation score was averaged over all SNR levels. Table II, shows the comparison of the TFM-separated signal with the proposed method, as well as the clean spectral phase scenario. Three tests are conducted for two TFMs, i.e., estimated BM and estimated RM, resulting in six paired tests. Within the three tests we assess the significance of the following:

- TFM (mixture phase) versus TFM (enhanced phase) as proposed.
- TFM (mixture phase) versus TFM (clean phase) as upper-bound.
- TFM (enhanced phase) as proposed versus TFM (clean phase) as upper-bound.

The Kolmogorov-Smirnov (KS) test rejects the null hypotheses for a p-value lower than 0.05, which means that the two compared tests are significantly different. The confidence intervals and p-values as outcomes of KS tests are also shown in the columns of the Table II. Except for STOI (comparing the TFM enhanced phase as proposed versus the TFM-separated using mixed signal phase), the KS test rejects the null hypotheses for all evaluation scores, concluding the significance of the reported results in the current section.

2. Experiment two: Speech mixture in noise

Finally, we report the results of the second experiment averaged on the whole utterances and noise types at different SNRs. To exemplify the effectiveness of such a challenging scenario, Fig. 6 and Fig. 7 show the bar plots of the proposed phase-enhanced TFM compared to the NMF algorithm proposed in Virtanen *et al.* (2013). To justify that phase modification is helpful for various source separation methods known in literature, we chose NMF as model-based, and IBM/IRM as CASA-driven methods. Similar to Mayer and Mowlaee (2015), we also provide the phase enhanced results for the NMF algorithm. For the sake of simplicity we averaged the results for target and masker with their appropriate SSR, to rivet on the separation performance only.

The following observations are made:

TABLE II. Significance test results, ascertained by computing a two-paired Kolmogorov-Smirnov test, observing the significant difference between TFM-separated signals and signals improved by the use of a proper phase estimate. The test rejects the null hypothesis for p-values lower than 0.05, indicating that both methods are significantly different.

	PESQ		STOI	
Compared Methods	Conf. Interval	p-Value	Conf. Interval	p-Value
est. BM (enh. phase) vs. est. BM (mixture phase)	2.16 ± 0.016 2.04 ± 0.015	$p < 0.05$	0.792 ± 0.0055 0.782 ± 0.0055	0.0537
est. BM (clean phase) vs. est. BM (mixture phase)	2.42 ± 0.01 2.04 ± 0.0152	$p < 0.05$	0.825 ± 0.00533 0.782 ± 0.0055	$p < 0.05$
est. BM (clean phase) vs. est. BM (enh. phase)	2.42 ± 0.0188 2.16 ± 0.0165	$p < 0.05$	0.825 ± 0.00533 0.792 ± 0.00552	$p < 0.05$
est. RM (enh. phase) vs. est. RM (mixture phase)	2.38 ± 0.0141 2.3 ± 0.0126	$p < 0.05$	0.807 ± 0.0052 0.798 ± 0.0053	0.109
est. RM (clean phase) vs. est. RM (mixture phase)	2.69 ± 0.0192 2.3 ± 0.0126	$p < 0.05$	0.849 ± 0.00477 0.798 ± 0.00525	$p < 0.05$
est. RM (clean phase) vs. est. RM (enh. phase)	2.69 ± 0.0192 2.38 ± 0.0141	$p < 0.05$	0.849 ± 0.00477 0.807 ± 0.0052	$p < 0.05$
	SDR		SIR	
Compared Methods	Conf. Interval	p-Value	Conf. Interval	p-Value
est. BM (enh. phase) vs. est. BM (mixture phase)	6.89 ± 0.0835 6.06 ± 0.0746	$p < 0.05$	11.7 ± 0.15 9.52 ± 0.129	$p < 0.05$
est. BM (clean phase) vs. est. BM (mixture phase)	11.7 ± 0.0955 6.06 ± 0.0746	$p < 0.05$	21.1 ± 0.109 9.52 ± 0.129	$p < 0.05$
est. BM (clean phase) vs. est. BM (enh. phase)	11.7 ± 0.0955 6.89 ± 0.0835	$p < 0.05$	21.1 ± 0.109 11.7 ± 0.15	$p < 0.05$
est. RM (enh. phase) vs. est. RM (mixture phase)	6.89 ± 0.0858 6.23 ± 0.08	$p < 0.05$	10.6 ± 0.135 8.57 ± 0.12	$p < 0.05$
est. RM (clean phase) vs. est. RM (mixture phase)	12.1 ± 0.0999 6.23 ± 0.08	$p < 0.05$	21.3 ± 0.112 8.57 ± 0.12	$p < 0.05$
est. RM (clean phase) vs. est. RM (enh. phase)	12.1 ± 0.0999 6.89 ± 0.0858	$p < 0.05$	21.3 ± 0.112 10.6 ± 0.135	$p < 0.05$

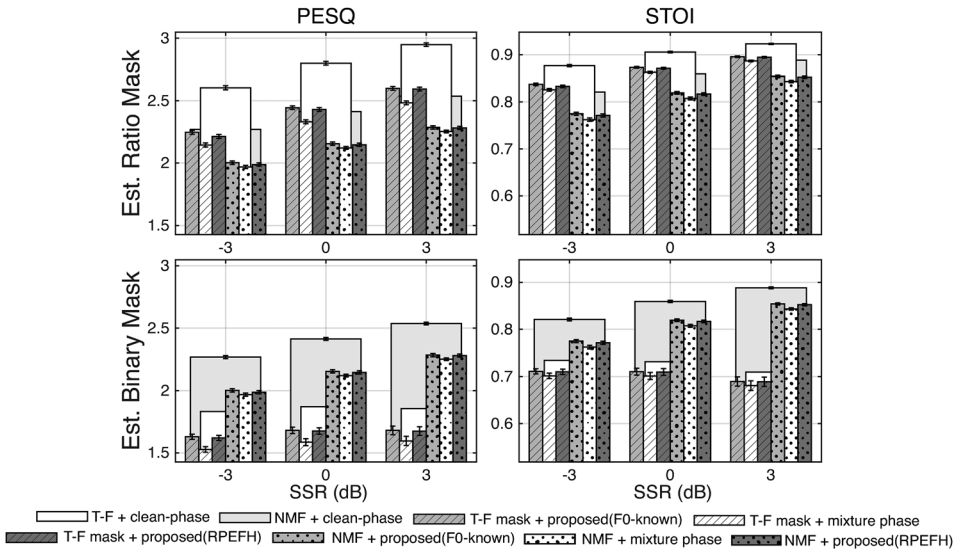

FIG. 6. Evaluation results: Estimated (top) IRM and (bottom) IBM compared to NMF in terms of quality (PESQ) and intelligibility (STOI). Results for oracle phase as well as IBM, IRM, and NMF (mixture phase) are reported for comparison.

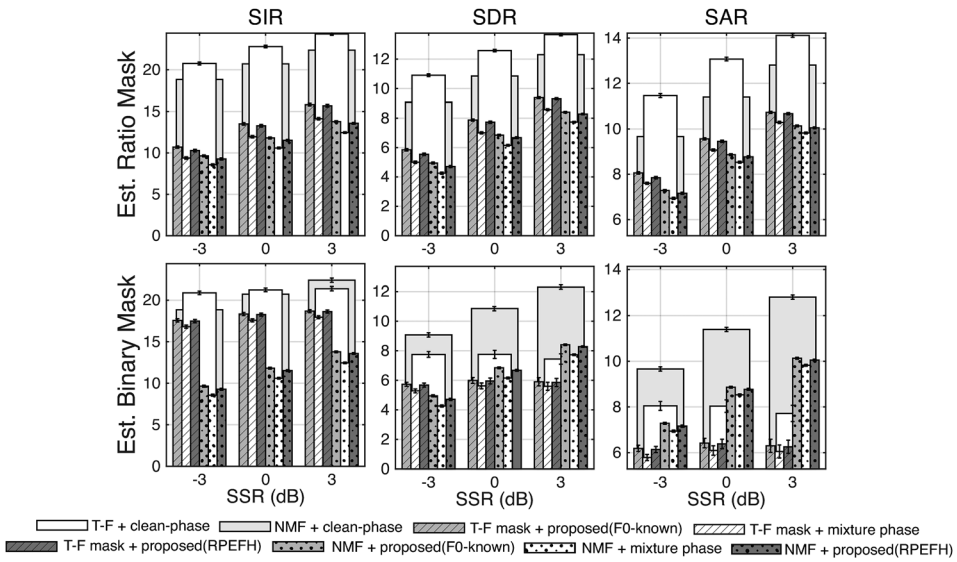

FIG. 7. Evaluation results: Estimated (top) IRM and (bottom) IBM in terms of separation performance (SIR), (SDR), and (SAR). Results for oracle phase as well as IBM, IRM, and NMF (mixture phase) are reported for comparison.

- In terms of PESQ and STOI, a consistent improvement for all SSRs is achieved when the estimated phase is used for the estimated IBM, IRM, and NMF rather than making use of the mixed signal phase. Similar to experiment one, the use of a proper phase estimate leads to an improvement of SIR and SDR. In terms of SAR the proposed algorithm produces less artifacts compared to its mixture phase counterpart.
- The two upper-bounds emphasize the achievable separation performance and highlight the importance of phase-aware processing for time-frequency mask method.
- We emphasize that a proper phase estimate is capable of removing noise outliers from the estimated IRM, albeit the estimated IBM eliminates the misclassified T-F cells which cannot be restored. This explains why estimated IRM performs better than estimated IBM when combined with phase enhancement. Nevertheless, in both time-frequency mask schemes, the proposed phase enhancement method improves the amplitude indirectly because of the constructive overlap-add routine during signal reconstruction.

VI. CONCLUSION

The conventional single-channel source separation methods mostly apply the phase of the mixed signal at signal reconstruction stage. For spectral amplitude enhancement, time-frequency masking (TFM) is often applied in either a binary or soft way. In the current contribution, we proposed an estimator for the clean spectral phase to replace the phase of the mixed signal at signal reconstruction. The enhanced spectral phase is combined with the estimated spectral amplitude provided by a time-frequency mask estimated in a blind scenario. The method relies on the fundamental frequency trajectory of the target source which is provided by a proposed histogram-based pitch estimator. Given this fundamental frequency estimate, the linear phase part is removed from harmonic phase to obtain an unwrapped harmonic phase. By applying temporal smoothing an enhanced spectral phase is obtained for signal reconstruction stage where the enhanced phase is combined with the TFM estimate for the spectral amplitude. It is important to note, that at high harmonics (overtones), less smoothing length or no smoothing is appropriate. At low harmonics, the algorithm benefits from a larger smoothing length. Therefore, an adaptive smoothing length should be considered for future works.

Experiments were conducted to evaluate the effectiveness of the proposed method when applied on estimated TFM as binary and ratio masks in a blind scenario. Comparing the proposed method with clean phase, mixture phase and benchmark phase enhancement methods showed that the proposed method was capable of better retrieving the desired harmonic structure of the target signal. Also, the proposed method led to joint improved perceived quality and speech intelligibility predicted by PESQ and STOI, and source separation outcome predicted by BSS EVAL measures. For ideal mask upper-bounds considerable gap between estimated and clean phase was observed, even if the amplitude is ideally estimated. We further point out that this gap motivates for further study on proposing novel phase estimator from mixed signals to achieve a performance closer to the clean phase reconstructed TFM upper-bound. While the current study was dedicated to single-channel speech separation, future work could be dedicated to apply the developed phase estimation techniques in audio source separation frameworks, like in Salamon *et al.* (2014), to improve the estimation of the magnitude and phase spectra of the underlying sources in the mixture.

ACKNOWLEDGMENTS

The work of F.M. and P.M. was supported by the Austrian Science Fund (project number P28070-N33). The work of D.S.W. and D.W. was supported in part by an ASFOR grant (FA 9550-12-1-0130), an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

[1]The prior work (Mayer and Mowlaee, 2015) used phase estimation combined with ideal binary mask to demonstrate the achievable upper-bound performance.

[2]Audio files are available at http://www2.spsc.tugraz.at/people/pmowlaee/PEISCSSTFM.html.

Babacan, O., Drugman, T., d'Alessandro, N., Henrich, N., and Dutoit, T. (**2013**). "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, pp. 7815–7819.

Bronson, J., and Depalle, P. (**2014**). "Phase constrained complex NMF: Separating overlapping partials in mixtures of harmonic musical sources," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp. 7475–7479.

Chu, W., and Alwan, A. (**2009**). "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, pp. 3969–3977.

Degottex, G., and Erro, D. (**2014**). "A uniform phase representation for the harmonic model in speech synthesis applications," EURASIP J. Audio Speech Music Processing **2014**, 38.

Duchi, J., Hazan, E., and Singer, Y. (**2011**). "Adaptive subgradient methods for online learning and stochastic optimization," J. Mach. Learn. Res. **12**, 2121–2159.

Févotte, C., and Godsill, S. J. (**2005**). "A Bayesian approach to time-frequency based blind source separation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk, NY.

Gerkmann, T. (**2014**). "MMSE-optimal enhancement of complex speech coefficents with uncertain prior knowledge of the clean speech phase," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp. 4511–4515.

Gerkmann, T., and Krawczyk, M. (**2013**). "MMSE-optimal spectral amplitude estimation given the STFT-phase," IEEE Signal Processing Lett. **20**(2), 129–132.

Gerkmann, T., Krawczyk-Becker, M., and Le Roux, J. (**2015**). "Phase processing for single-channel speech enhancement: History and recent advances," IEEE Signal Processing Mag. **32**(2), 55–66.

Glorot, X., Bordes, A., and Bengio, Y. (**2011**). "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, Fort Lauderdale, FL, edited by G. J. Gordon and D. B. Dunson, Vol. 15, pp. 315–323.

Gonzalez, S., and Brookes, M. (**2014**). "PEFAC - a pitch estimation algorithm robust to high levels of noise," IEEE Trans. Audio Speech Language Processing **22**(2), 518–530.

Griffin, D., and Lim, J. (**1984**). "Signal estimation from modified short-time Fourier transform," IEEE Trans. Acoust. Speech Signal Processing **32**(2), 236–243.

Gunawan, D., and Sen, D. (**2010**). "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," IEEE Signal Processing Lett. **17**(5), 421–424.
Harris, F. J. (**1978**). "On the use of windows for harmonic analysis with the discrete Fourier transform," Proc. IEEE **66**(1), 51–83.
Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**(4), 3029–3038.
Hershey, J. R., Rennie, S. J., Olsen, P. A., and Kristjansson, T. T. (**2010**). "Super-human multi-talker speech recognition: A graphical modeling approach," Comput. Speech Language **24**(1), 45–66.
IEEE Audio and Electroacoustics Group (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**(3), 225–246.
ITU Radiocommunication Assembly (**2001**). "ITU-T P. 862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Technical report, ITU, Geneva, Switzerland.
Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**(3), 1486–1494.
Koutsogiannaki, M., Simantiraki, O., Degottex, G., and Stylianou, Y. (**2014**). "The importance of phase on voice quality assessment," in *15th Annual Conference of the International Speech Communication Association (ISCA)*, Singapore, pp. 1653–1657.
Kulmer, J., and Mowlaee, P. (**2014**). "Phase estimation in single channel speech enhancement using phase decomposition," IEEE Signal Processing Lett. **22**(5), 598–602.
Le Roux, J., and Vincent, E. (**2013**). "Consistent Wiener filtering for audio source separation," IEEE Signal Processing Lett. **20**, 217–220.
Loizou, P. C. (**2013**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL).
Magron, P., Badeau, B., and David, B. (**2015a**). "Phase recovery in NMF for audio source separation: An insightful benchmark," in *International Conference on Acoustics Speech, Signal Processing (ICASSP)*, Brisbane, Australia, pp. 81–85.
Magron, P., Badeau, B., and David, B. (**2015b**). "Phase reconstruction of spectrograms with linear unwrapping: Application to audio signal restoration," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Nice, France, pp. 1–5.
Mayer, F., and Mowlaee, P. (**2015**). "Improved phase reconstruction in single-channel speech separation," in *16th Annual Conference of the International Speech Communication Association (ISCA)*, Dresden, Germany, pp. 1795–1799.
Mowlaee, P. (**2010**). "New stategies for single-channel speech separation," Ph.D. thesis, Institut for Elektroniske Systemer, Aalborg Universitet, Aalborg, Denmark.
Mowlaee, P., and Kulmer, J. (**2015a**). "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and snr information," IEEE Trans. Audio Speech Language Processing **23**(9), 1521–1532.
Mowlaee, P., and Kulmer, J. (**2015b**). "Phase estimation in single-channel speech enhancement: Limits-potential," IEEE Trans. Audio Speech Language Processing **23**(8), 1283–1294.
Mowlaee, P., Kulmer, J., Stahl, J., and Mayer, F. (**2016a**). *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice* (Wiley, New York), 256 pp.
Mowlaee, P., and Saeidi, R. (**2014**). "Time-frequency constraint for phase estimation in single-channel speech enhancement," in *Proceedings International Workshop on Acoustic Signal Enhancement*, pp. 338–342.
Mowlaee, P., Saeidi, R., Christensen, M. G., and Martin, R. (**2012b**). "Subjective and objective quality assessment of single-channel speech separation algorithms," in *Proceedings of the IEEE International Conference of Acoustics, Speech, Signal Processing (ICASSP)*, Kyoto, Japan, pp. 69–72.
Mowlaee, P., Saeidi, R., and Martin, R. (**2012a**). "Phase estimation for signal reconstruction in single-channel speech separation," in *13th Annual Conference of the International Speech Communication Association (ISCA)*, Portland, OR, pp. 1548–1551.
Mowlaee, P., Saeidi, R., and Stylianou, Y. (**2014**). "Phase importance in speech processing applications," in *15th Annual Conference of the International Speech Communication Association (ISCA)*, Singapore, pp. 1623–1627.
Mowlaee, P., Saeidi, R., and Stylianou, Y. (**2016b**). "Advances in phase-aware signal processing for speech communication," Speech Commun. **81**, 1–29.
Salamon, J., Gómez, E., Ellis, D., and Richard, G. (**2014**). "Melody extraction from polyphonic music signals: Approaches, applications and challenges," IEEE Signal Processing Mag. **31**(2), 118–134.
Sturmel, N., and Daudet, L. (**2012**). "Iterative phase reconstruction of Wiener filtered signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, pp. 101–104.
Sturmel, N., and Daudet, L. (**2013**). "Informed source separation using iterative reconstruction," IEEE Trans. Audio Speech Language Processing **21**(1), 178–185.
Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Language Processing **19**(7), 2125–2136.
Vary, P. (**1985**). "Noise suppression by spectral magnitude estimation mechanism and theoretical limits," Signal Processing **8**(4), 387–400.
Vincent, E., Gribonval, R., and Févotte, C. C. (**2006**). "Performance measurement in blind audio source separation," IEEE Trans. Audio Speech Language Processing **14**(4), 1462–1469.
Virtanen, T., Gemmeke, J. F., and Raj, B. (**2013**). "Active-set Newton algorithm for overcomplete non negative representations of audio," IEEE Trans. Audio Speech Language Processing **21**(11), 2277–2289.
Virtanen, T., Gemmeke, J. F., Raj, B., and Smaragdis, P. (**2015**). "Compositional models for audio processing: Uncovering the structure of sound mixtures," IEEE Signal Processing Mag. **32**(2), 125–144.
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Springer, New York), pp. 181–197.
Wang, D., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, New York), 395 p.
Wang, D., and Lim, J. S. (**1982**). "The unimportance of phase in speech enhancement," IEEE Trans. Audio Speech Language Processing **30**(4), 679–681.
Wang, Y., Han, K., and Wang, D. (**2013**). "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio Speech Language Processing **21**(2), 270–279.
Wang, Y., Narayanan, A., and Wang, D. (**2014**). "On training targets for supervised speech separation," IEEE Trans. Audio Speech Language Processing **22**(12), 1849–1858.
Williamson, D. S., Wang, Y., and Wang, D. (**2014**). "Reconstruction techniques for improving the perceptual quality of binary masked speech," J. Acoust. Soc. Am. **136**(2), 892–902.
Williamson, D. S., Wang, Y., and Wang, D. (**2015**). "Estimating non-negative matrix model activations with deep neural networks to increase perceptual speech quality," J. Acoust. Soc. Am. **138**, 1399–1407.
Zhang, X., and Wang, D. (**2014**). "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *15th Annual Conference of the International Speech Communication Association (ISCA)*, Singapore, pp. 1534–1538.