On Optimal Filtering for Speech Decomposition

Alfredo Esquivel Jaramillo, Jesper Kjær Nielsen and Mads Græsbøll Christensen
Audio Analysis Lab CREATE, Aalborg University
Emails: {aeja, jkn,mgc}@create.aau.dk

***Abstract*—Optimal linear filtering has been used extensively for speech enhancement. In this paper, we take a first step in trying to apply linear filtering to the decomposition of a noisy speech signal into its components. The problem of decomposing speech into its voiced and unvoiced components is considered as an estimation problem. Assuming a harmonic model for the voiced speech, we propose a Wiener filtering scheme which estimates both components separately in the presence of noise. It is shown under which conditions this optimal filtering formulation outperforms two state-of-the-art speech decomposition methods, which is also revealed by objective measures, spectrograms and informal listening tests.**

***Index Terms*—Speech decomposition, time-domain filtering, Wiener filter, voiced speech, unvoiced speech.**

I. Introduction

The decomposition of speech into its major components, i.e., voiced and unvoiced, is a challenge in many speech processing applications. An accurate recovery of these components is important in speech coding [1], analysis [2], synthesis [3], enhancement [4], as well as for diagnosing illnesses [5]. The presence of noise is inevitable in most acoustic scenarios, so a major challenging problem is the robust estimation of both components in the presence of additive noise. This is useful, for example, in remote voice assessment applications [6], and therefore, for a proper diagnosis of voice pathologies. Many clinical assesment systems have used sustained vowel phonations to detect voice pathologies, and recently [7] the need of assessing natural speech has been considered.

With the classical speech production model, speech is classified as voiced or unvoiced depending on whether the source is a periodic impulse or a white noise sequence [8]. However, specially for a good quality of synthetic speech, it has been shown [9] that a mixed excitation can produce a more natural sounding speech. This is the case for voiced fricatives (e.g. /z/). Additionally, for clinical assessment of voice impairment, it is necessary to take into account the presence of the white noise source (i.e. unvoiced component) in the vocal apparatus which results in breathy vowels and other forms of vocal dysphonia [10].

Some efforts to separate the voiced and unvoiced components from a speech signal have been developed. There are methods which make a binary voiced/unvoiced decision per frequency bin such as the one based on the multiband excitation vocoder [1] and the harmonic plus noise (HNS) model [11], and methods [12], [13] which consider both components can coexist in the speech frequency bands, which is more accurate from a speech production perspective [12]. The well-known methods at this respect are an iterative method for a periodic and aperiodic excitation decomposition [12] and a pitch scaled harmonic filtering (PSHF) based method [13]. The iterative method operates on an assumed mixed excitation to the vocal tract by reconstructing the unvoiced part excitation in the harmonic regions which are obtained from the cepstrum. The PSHF method is based on a pitch-scaled least-squares separation of the speech signal in the frequency domain. These speech decomposition methods, which decompose speech signals into stochastic and deterministic components, do not take the presence of background noise into account in the decomposition, and thus do not distinguish between and deal with unvoiced speech and noise, which may be present at the same time.

In the speech enhancement literature, a common approach to estimate a clean signal corrupted by noise is optimal filtering, such as the classical Wiener filter [14]. Traditionally, the filter design requires estimates of the second-order statistics of the noisy signal and the noise. In this paper, we investigate if the speech decomposition problem can also be tackled via an optimal filtering way. To use optimal filtering for decomposing speech into its components, we need estimates of their second-order statistics. To obtain these, we assume a periodic signal model, namely the harmonic model, for the voiced component [15], [16]. By assuming stationarity in a short time segment, the statistics of the voiced component will depend on the fundamental frequency, the number of harmonics and the power of the harmonics. If the noise is stationary, its statistics can be estimated during periods where no voice activity is detected. Otherwise, they can be obtained through the principle of minimum tracking [17], for example. Knowing the statistics of the voiced part, of the noise and those of the observed signal, the statistics of the unvoiced part can be estimated, and, therefore, a Wiener filter can be employed to extract separately the voiced and unvoiced component.

The remainder of the paper is structured as follows. Section II introduces the signal model, the assumptions and the optimal filtering formulation. Section III establishes the proposed filtering approach and details the main parts of the statistics estimation for each of the components. Section IV gives the performance measures and the experimental results. Finally, the paper is concluded in section V.

This work is funded by Independent Research Fund, Denmark and by the National Council of Science and Technology, CONACYT, Mexico, under the grant 418437.

ISBN 978-90-827970-1-5 © EURASIP 2018

II. Signal model and Problem formulation

The speech decomposition problem considered in this paper is to extract both the zero-mean voiced $v(n)$ and unvoiced $u(n)$ components, from the noisy observation $y(n)$, i.e,

$$y(n) = s(n) + z(n) = v(n) + u(n) + z(n), \tag{1}$$

where $n = 0, 1, \ldots, N-1$ is the discrete-time index, $s(n) = v(n) + u(n)$ is the clean speech signal which is buried in a zero-mean additive white or colored noise $z(n)$. We assume that the voiced and unvoiced parts as well as the noise are uncorrelated.

When we adopt a linear filtering approach to recovering the desired speech components, we consider the M recent successive samples. Therefore, the signal model in (1) can be expressed in a vector form as

$$\mathbf{y}(n) = \mathbf{v}(n) + \mathbf{u}(n) + \mathbf{z}(n), \tag{2}$$

where $\mathbf{y}(n) = [y(n)\ y(n-1)\ \ldots\ y(n-M+1)]^T$ is a vector of length M, $[\cdot]^T$ denotes the transpose of a vector or matrix, and $\mathbf{v}(n)$, $\mathbf{u}(n)$ and $\mathbf{z}(n)$ are defined in a similar way to $\mathbf{y}(n)$. The objective of speech decomposition is to estimate one or more samples of $v(n)$ and $u(n)$ from the noisy vector $\mathbf{y}(n)$ by the application of two different optimal filters to the observed signal vector, i.e.

$$\begin{aligned}\hat{v}(n) &= \sum_{k=0}^{M-1} h_{v,k} y(n-k) \\ &= \mathbf{h}_v^T \mathbf{y}(n) = \mathbf{h}_v^T \mathbf{v}(n) + \mathbf{h}_v^T \mathbf{u}(n) + \mathbf{h}_v^T \mathbf{z}(n),\end{aligned} \tag{3}$$

$$\hat{u}(n) = \mathbf{h}_u^T \mathbf{y}(n) = \mathbf{h}_u^T \mathbf{u}(n) + \mathbf{h}_u^T \mathbf{v}(n) + \mathbf{h}_u^T \mathbf{z}(n) \tag{4}$$

where $\mathbf{h}_v = [h_{v,0}\ \ldots\ h_{v,M-1}]^T$, $\mathbf{h}_u = [h_{u,0}\ \ldots\ h_{u,M-1}]^T$, and $\hat{v}(n)$, $\hat{u}(n)$ are estimates of $v(n)$ and $u(n)$ respectively.

For speech decomposition, the problem is to find the optimal filters $\mathbf{h}_v$ and $\mathbf{h}_u$ which make the level of the undesired components as small as possible while passing the desired component with as little distortion as possible. The undesired components are the sum of the two last right-hand terms of (3) and (4). With the assumption that $v(n)$, $u(n)$ and $z(n)$ are uncorrelated, the $M \times M$ covariance matrix of the observed signal can be expressed as

$$\mathbf{R}_\mathbf{y} = E\left[\mathbf{y}(n)\mathbf{y}^T(n)\right] = \mathbf{R}_\mathbf{v} + \mathbf{R}_\mathbf{u} + \mathbf{R}_\mathbf{z}, \tag{5}$$

where $E[\cdot]$ denotes expectation, $\mathbf{R}_\mathbf{v} = E\left[\mathbf{v}(n)\mathbf{v}^T(n)\right]$, $\mathbf{R}_\mathbf{u} = E\left[\mathbf{u}(n)\mathbf{u}^T(n)\right]$, $\mathbf{R}_\mathbf{z} = E\left[\mathbf{z}(n)\mathbf{z}^T(n)\right]$ are the covariance matrices of $\mathbf{v}(n)$, $\mathbf{u}(n)$, and $\mathbf{z}(n)$, respectively.

III. Optimal Filtering and Statistics Estimation

By considering the error between the true voiced and the estimated voiced component, i.e., $e_v(n) = \mathbf{h}_v^T \mathbf{y}(n) - v(n)$, and the error between the true unvoiced and the estimated unvoiced component, i.e. $e_u(n) = \mathbf{h}_u^T \mathbf{y}(n) - u(n)$, the mean-squared-error (MSE) criteria can be defined as

$$J_v(\mathbf{h}_v) = E\left[e_v^2(n)\right] = \sigma_v^2 - 2\mathbf{h}_v^T \mathbf{R}_v \mathbf{i}_1 + \mathbf{h}_v^T \mathbf{R}_y \mathbf{h}_v, \tag{6}$$

$$J_u(\mathbf{h}_u) = E\left[e_u^2(n)\right] = \sigma_u^2 - 2\mathbf{h}_u^T \mathbf{R}_u \mathbf{i}_1 + \mathbf{h}_u^T \mathbf{R}_y \mathbf{h}_u \tag{7}$$

where $\mathbf{i}_1$ is the first column of the $M \times M$ identity matrix $\mathbf{I}_M$, and σ_v^2 and σ_u^2 are the variances of $v(n)$ and $u(n)$, respectively. If we take the gradient of each MSE with respect to $\mathbf{h}_v$ and $\mathbf{h}_u$, and equate the results to $\mathbf{0}$, we find the Wiener filters for estimating the voiced and unvoiced speech components to

$$\mathbf{h}_v = \mathbf{R}_y^{-1} \mathbf{R}_v \mathbf{i}_1, \tag{8}$$

$$\mathbf{h}_u = \mathbf{R}_y^{-1} \mathbf{R}_u \mathbf{i}_1. \tag{9}$$

To compute these filters, the different statistics in (5) are required. In order to avoid problems over frame transitions of the noisy signal, we adopt a recursive approach [18], in which a short-term sample estimate and a moving average is used for computing an estimate at the time frame n as

$$\hat{\mathbf{R}}_y(n) = \alpha_y \hat{\mathbf{R}}_y(n-1) + (1-\alpha_y)\bar{\mathbf{R}}_y(n), \tag{10}$$

where $0 < \alpha_y < 1$ is a forgetting factor and

$$\bar{\mathbf{R}}_y(n) = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{y}(n)\mathbf{y}^H(n). \tag{11}$$

For the voiced part $v(n)$, we use the harmonic model, i.e.

$$v(n) = \sum_{l=1}^{L} A_l \cos(l\omega_0 n + \phi_l), \tag{12}$$

where L is the number of harmonics, ω_0 is the fundamental frequency, A_l denotes the real amplitude of the lth harmonic with its corresponding phase $\phi_l \in [0, 2\pi)$. As an extension to the vector model in (2), the voiced signal vector is expressed as $\mathbf{v}(n) = \mathbf{Z}\mathbf{a}$, with the definitions

$$\mathbf{a} = \frac{1}{2}[A_1 e^{j\phi_1}\ A_1 e^{-j\phi_1}\ \ldots\ A_L e^{j\phi_L}\ A_L e^{-j\phi_L}]^T, \tag{13}$$

$$\mathbf{Z} = [\mathbf{z}(\omega_0)\, \mathbf{z}^*(\omega_0)\ \ldots\ \mathbf{z}(\omega_0 L)\ \mathbf{z}^*(\omega_0 L)], \tag{14}$$

$$\mathbf{z}(\omega_0 l) = [1\ e^{jl\omega_0} \ldots e^{jl(M-1)\omega_0}]^T. \tag{15}$$

The voiced part covariance matrix $\mathbf{R}_\mathbf{v} = E\{\mathbf{v}(n)\mathbf{v}^H(n)\} = E\{(\mathbf{Z}\mathbf{a})(\mathbf{Z}\mathbf{a})^H\}$ can be expressed as $\mathbf{R}_\mathbf{v} \approx \mathbf{Z}\mathbf{P}\mathbf{Z}^H$ [19], where $[\cdot]^H$ denotes complex conjugate transpose and the amplitude covariance matrix $\mathbf{P}$ has the form [19]

$$\mathbf{P} = E\{\mathbf{a}\mathbf{a}^H\} = \frac{1}{4}\,\mathrm{diag}([A_1^2\, A_1^2 \ldots\, A_L^2\, A_L^2]). \tag{16}$$

Clearly, $\mathbf{R}_\mathbf{v}$ depends on ω_0, the model order L and the amplitude vector $\mathbf{a}$, which need to be estimated. The amplitude vector can be estimated using the principle of least-squares [20] as $\hat{\mathbf{a}} = (\mathbf{Z}^H\mathbf{Z})^{-1}\mathbf{Z}^H\mathbf{y}$, and the fundamental frequency and model order L are estimated by using a fast nonlinear least squares (NLS) algorithm [21]. However, the NLS method assumes that the signal is observed in white gaussian noise, which is not always true in many real acoustic cases. Therefore, after estimating the noise power spectral density, a linear prediction scheme suggested in [22] is used to prewhiten the noisy signal. Then, the fundamental frequency is estimated from the prewhitened signal resulting in better

ISBN 978-90-827970-1-5 © EURASIP 2018

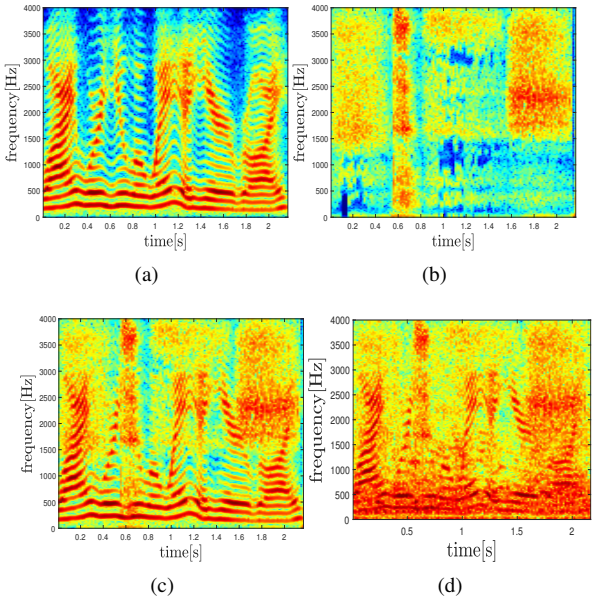

Figure 1: Spectrograms of (a) the true voiced component, (b) the true unvoiced component (concatenation of sounds in the order /f/, /t/, /s/, /sh/, (c) the clean speech (true voiced plus true unvoiced), and (d) the noisy speech with iSNR = 4 dB.

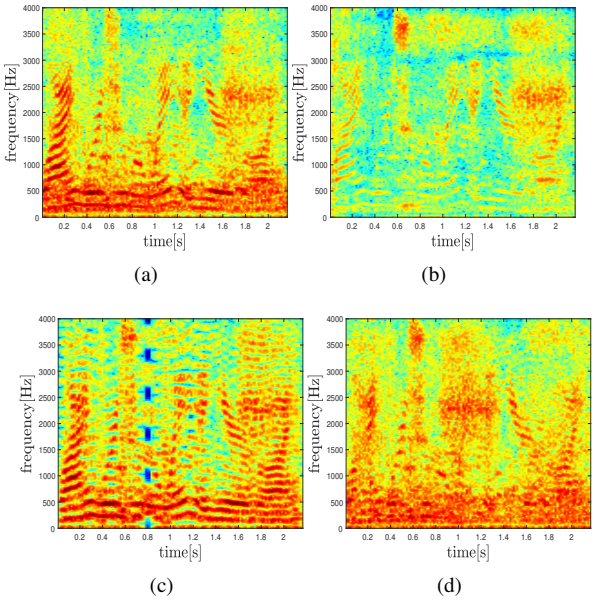

Figure 2: Spectrograms of (a) the estimated voiced component using the proposed filtering approach, (b) the estimated unvoiced component using the proposed filtering approach, (c) the voiced component obtained by ITER algorithm, and (d) the unvoiced component by PSHF method.

frequency estimates than without prewhitening, when dealing with speech corrupted in colored noise.

As voiced speech is non-stationary across a segment of length N, a similar recursive approach to (10) can be used to smooth the voiced frame covariance matrix

$$\hat{\mathbf{R}}_v(n) = \alpha_v \hat{\mathbf{R}}_v(n-1) + (1-\alpha_v)\bar{\mathbf{R}}_v(n), \qquad (17)$$

where $\bar{\mathbf{R}}_v(n) = \mathbf{Z}\mathbf{P}\mathbf{Z}^H$ and $0 < \alpha_v < 1$ is another forgetting factor. A noise estimator based on optimal smoothing and minimum statistics [17], for example, can be used to estimate $\mathbf{R}_z$. From (5), after the voiced part and noise covariance matrices are estimated, an estimate of the unvoiced component covariance matrix at the time frame n can be computed as $\hat{\mathbf{R}}_u(n) = \hat{\mathbf{R}}_y(n) - \hat{\mathbf{R}}_v(n) - \hat{\mathbf{R}}_z(n)$. To ensure that this matrix is positive definite, an eigenvalue decomposition is applied and its negative eigenvalues are replaced with a very small positive number [23].

IV. Experimental Results.

In this section, the performance of the proposed filtering approach (optimal) is compared to the iterative periodic-aperiodic decomposition (ITER) [12] and the pitch scaled harmonic filter (PSHF) [13] based method for noisy speech signal decomposition. The state-of-the-art methods evaluated their performance in a quantitative way only for synthetic speech signals, since the individual speech components are not available separately for real speech [24]. As it is difficult to evaluate the quality of a given decomposition in an objective way, we consider an intermediate approach, where we mix fully voiced and fully unvoiced utterances, so that we know the ground truth components of speech. The mixing of the two signals may not sound as natural as one would expect for common speech, but it will allow us to compare the decomposition performance with objective measures.

In the experiments, we consider five fully voiced utterances [25] (4 male and 1 female) as a ground truth for the voiced component, resampled to a sampling frequency of 8 kHz. In Fig. 1(a), the spectrogram of the fully voiced female utterance "Why were you away a year, Roy?" is shown. For the unvoiced speech component, we consider the concatenation of five sounds /sh/, /f/, /s/, /t/, /p/ from the audio recordings of a free ebook about the full range of sounds used in general British English pronunciation [26], also resampled to 8 kHz. These sounds are either unvoiced fricatives or unvoiced stops [8]. As can be seen from Fig. 1(b), this recording does lack a harmonic structure and has the appearance of rectangular red patterns instead of horizontal striations [8], which is representative of unvoiced speech. The clean speech for the experiment is the sum of the voiced speech and unvoiced speech, where different combinations of the five voiced sentences and ten orderings of the unvoiced sounds are considered, and the results will be averaged across the different realizations. An example of a clean signal, which contains both voiced and unvoiced parts, is shown in Fig. 1(c). Three types of noise are considered: white, street and babble. The recordings of the street and babble noises are taken from the AURORA database [27]. In Fig. 1(d), is shown the noisy speech spectrogram, which is formed by adding babble noise to the clean speech, the input SNR is 4 dB.

ISBN 978-90-827970-1-5 © EURASIP 2018

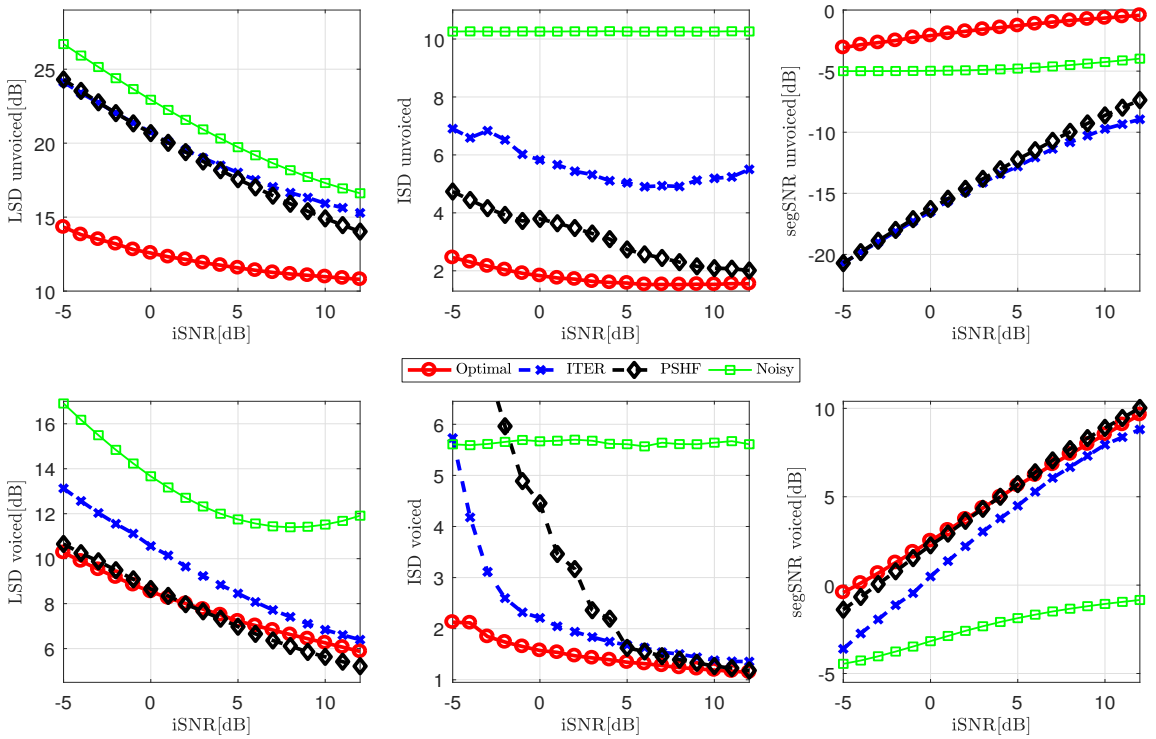

Figure 3: Average measured LogSpectrum distortion (LSD), Itakura-Saito distance (ISD) and segmental SNR(segSNR) for the proposed approach and state-of-the-art methods in different iSNRs and noise types. Comparison also includes the case of noisy speech as an estimate.

For the comparison, we add different types of noise to the speech signal at certain iSNR, ranging from -5 dB to 12 dB, two different noisy realizations at each iSNR are considered for each possible combination of voiced and unvoiced speech. For the proposed approach, the segment length is set to $N = 200$ and the filter length to $M = 25$, the forgetting factors to $\alpha_y = \alpha_v = 0.75$ in the white noise scenario and $\alpha_y = \alpha_v = 0.96$ in the street and babble noise scenario. The noise statistics are estimated using the minimum statistics (MS) [17] principle.

The spectrograms of the voiced and unvoiced component obtained by optimal filtering principle application, for the case of speech in babble noise (Fig. 1(d)), are shown in Fig. 2(a) and (b), the voiced component obtained by the ITER algorithm in Fig. 2(c) and the unvoiced component which results from the PSHF method in Fig. 2(d). The spectrogram in Fig. 2(b) shows that the herein developed approach generates an unvoiced estimated component which looks more similar to the original unvoiced speech signal (opposed to that of Fig. 2(d) in the sense that its spectrogram has similar red patterns as opposed to the PSHF method, which looks distributed in other frequency bins. Similar observations can be made with the ITER method. Even if some frequency bins do not appear in the spectrogram, informal listening tests reveal that the unvoiced stops or sounds can be perceived, so the main features of the unvoiced component are preserved.

The decomposition performance was evaluated quantitatively in terms of segmental SNR (segSNR), Itakura-Saito distance (ISD), [28] and LogSpectrum Distortion (LSD) [29]. Due to space constraint, we here show the result at each iSNR averaged across the different realizations and across all the three noise types. The results are plotted in Fig. 3. The comparison also includes the case of the noisy speech as an estimate, in order to see if the methods perform better or worse than the case of no processing of the noisy speech at all. Next, we describe what can be observed from the different plots of Fig. 3.

The presented approach not only outperforms the other two in terms of segSNR for unvoiced speech, but it also results in a better measure against the case in which the noisy speech is considered as an estimate. This does not happen for the other methods, whose performance is below the curves of noisy speech as an estimate. In fact, as can be seen from Fig. 2(d), the other methods show low-frequency content which is not present in the true unvoiced speech component, and that results in more signal content than this ground truth. The informal listening of their outputs does not allow to perceive all the unvoiced sounds, and some remaining of the female sentence with a high level of distortion can be listened in these unvoiced estimates. This does not occur by decomposing speech with the optimal filtering approach, in which the different unvoiced fricative and stop sounds can be perceived. In the white noise case for the ITER method, all the phonemes are lost, and for the PSHF method, only one of the phonemes is preserved, but in a very distorted manner. Much lower values of LogSpectrum distortion (LSD) and lower Itakura-Saito (ISD) distance values are also obtained with the optimal filtering formulation.

In the voiced speech case, the optimal filtering approach results in higher segSNR than the ITER method, and similar values with respect to the PSHF method at all iSNRs. It is

ISBN 978-90-827970-1-5 © EURASIP 2018

important to mention that babble noise is one of the most difficult noise types to remove, since it is highly nonstationary and contains similar spectral content to speech. In this paper, we considered the noise statistics estimated with the default settings of the minimum statistics approach [17], but in a future improvement, the developed principle herein can be combined with a codebook-based approach [30], in order to get better estimates of the noise statistics. With respect to the Itakura-Saito distance (ISD), the ISD of the voiced component obtained by the optimal filtering formulation is lower than the other methods. This measure is more perceptually relevant than the segSNR [28]. The spectrogram of the voiced component processed by the ITER algorithm reveals some higher frequency components (>3000Hz), which were not present in the true voiced speech, and also some harmonics below this frequency range, which were not present in the original speech. Informal listening test reveals that the voiced output of the ITER algorithm sounds more artificially distorted than the one obtained from the optimal filtering principle. For the developed approach, although the voiced estimate (Fig. 2(a)) has still some noise present, it preserves the original features of the ground truth and sounds less distorted than the other methods. Finally, with the optimal filtering decomposition approach, we observe similar LogSpectrum distortion (LSD) values to the PSHF method for all iSNRs, and the proposed approach also has lower LSD values than the ITER method. Even if LSD and segSNR are similar for both approaches (optimal and PSHF), the ISD of the voiced PSHF estimate is higher for low SNRs.

V. Conclusions.

In this paper, we have considered the speech decomposition problem employing the principle of optimal filtering with the corresponding statistics estimation for each one of the components of the noisy observation. We investigated if the presented approach is more robust and convenient for speech decomposition in noisy conditions. Based on the informal listening tests, spectrogram analysis and the objective measures, we found that the optimal filtering approach seems to work well.

References

[1] D. W. Griffin and J.S. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, aug 1988.

[2] P. Cook, "Noise and aperiodicity in the glottal source: A study of singer voices," in *Twelfth International Congress of Phoenetic Sciences*, 08/1991 1991, number STAN-M-75.

[3] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: improving production quality," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1870–1887, Dec 1989.

[4] J. Hardwick, C. D. Yoo, and J. S. Lim, "Speech enhancement using the dual excitation speech model," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1993, vol. 2, pp. 367–370 vol.2.

[5] Y. Qi, R. E. Hillman, and C. Milstein, "The estimation of signal-to-noise ratio in continuous speech for disordered voices," *The Journal of the Acoustical Society of America 105*, vol. 105, pp. 2532–2535, 1999.

[6] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, March 2006.

[7] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model," *Journal of voice*, vol. 30, pp. 757.e7–757.e19, 2016.

[8] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Wiley-IEEE Press, 2000.

[9] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," 1982, vol. 7, pp. 614–617, Institute of Electrical and Electronics Engineers.

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

[11] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, Jan 2001.

[12] B. Yegnanarayana, C. D'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, 1998.

[13] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, Oct 2001.

[14] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, July 2006.

[15] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 5969–5983, Dec 2010.

[16] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 731–740, 2001.

[17] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," vol. 9, no. 5, pp. 504–512, Jul. 2001.

[18] J. Chen, J. Benesty, and Y. A. Huang, "Study of the noise-reduction problem in the Karhunen-Loeve expansion domain," *IEEE Transactions on Speech and Audio Processing*, vol. 17, pp. 787–802, 2009.

[19] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2042–2056, Oct 2013.

[20] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: survey, new results, and an application," *IEEE Transactions on Signal Processing*, vol. 48, pp. 338–352, 2000.

[21] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Processing*, vol. 135, no. Supplement C, pp. 188 – 197, 2017.

[22] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for non-stationary voiced speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2354–2367, Dec 2016.

[23] J. Benesty and J. Chen, *A Conceptual Framework for Noise Reduction*, Springer, Germany, 2015.

[24] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 12–23, Jan 1998.

[25] M. Cooke, *Modelling Auditory Processing and Organisation*, Cambridge University Press, New York, NY, USA, 1993.

[26] P.S., *45 Sounds of GB English*, Pronunciation Studio, 2017.

[27] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[28] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[29] I. McLoughlin, *Applied Speech and Audio Processing: With Matlab Examples*, Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[30] J.K. Nielsen, M.S. Kavalekalam, M.G. Christensen, and J.B. Boldt, "Model-based noise psd estimation from speech in non-stationary noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

ISBN 978-90-827970-1-5 © EURASIP 2018