A POSTERIORI VOICED / UNVOICED PROBABILITY ESTIMATION BASED ON A SINUSOIDAL MODEL

Robert Rehr, Martin Krawczyk and Timo Gerkmann

University of Oldenburg, Germany, Cluster of Excellence Hearing4all,
Department of Medical Physics and Acoustics, Speech Signal Processing Group,
{firstname.lastname}@uni-oldenburg.de

ABSTRACT

In this paper, we focus on methods for estimating the a posteriori probability of a signal segment being voiced which employ a harmonic signal model. Fisher et al. [1] present two likelihood functions for voiced and unvoiced speech from which the posterior probability can be derived. However, due to the chosen models, the a posteriori probability of a signal segment being voiced does not go to 0 % in unvoiced speech. Thus, a novel algorithm is proposed, which incorporates the expected unvoiced speech energy and allows for obtaining low probabilities. Further, it explicitly models the statistics of the segment energy and employs a state-of-the-art noise tracker. Experiments which were conducted on the TIMIT database for different noise types and noise levels show that the proposed method results in lower over-estimation and under-estimation of the voicing probability as compared to [1].

Index Terms— voiced-unvoiced decision, likelihood ratio test, harmonic model, voicing determination, a posteriori probability

1. INTRODUCTION

In speech signals, two main excitation types of the human voice can be discriminated which are known as *voiced* and *unvoiced*. Voiced sounds are characterized by their periodic structure due to the vocal fold oscillation. In contrast, unvoiced segments reveal a noisy character which is the result of the air flow passing constrictions in the oral cavities. Both speech types are often treated differently in signal processing, e. g. in speech coding [2] or noise reduction applications [3], where the output may be a combination of two signals which were independently generated for the two excitation types. Here, the probability can be used for controlling a soft mixing of both signals and, as a consequence, it becomes an important parameter for these algorithms.

Different methods have been proposed for identifying voiced and unvoiced passages under noisy conditions. Early proposals employ features such as short-time signal energy or zero-crossing rate for distinguishing voiced and unvoiced segments, e. g. [4]. In [5] similar features are employed in an unsupervised learning scheme. Often, voicing determination algorithms (VDAs) are coupled to a fundamental frequency estimator. For example the methods proposed in [6, 7, 8] search for the cepstral coefficient that belongs to the fundamental frequency and compare its value with a threshold for making their decision on whether the signal is voiced or unvoiced. Also the VDAs proposed in [1, 9], which employ the harmonic model [10], are based on a fundamental frequency estimator. This model represents the speech signal as sinusoidal oscillations, whose frequencies are integer multiples of the fundamental frequency. Fisher et al. [1] use the model for constructing a VDA, where two likelihood functions for voiced and unvoiced speech are compared using a likelihood ratio test (LRT).

In this paper, we aim at obtaining the probability of a frame being voiced given a noisy observation, i. e. the a posteriori probability. For this, the likelihoods of [1] can be employed. However, it is possible to show that the likelihood models for voiced and unvoiced speech employed in [1] become identical in unvoiced speech periods. This prevents the discrimination between the voicing states and as a consequence, the a posteriori probability for the voiced hypothesis does not go to zero in unvoiced speech. Therefore, we propose a novel method which allows to obtain low probabilities in these segments. For this, we incorporate the energy of unvoiced speech in our unvoiced model in terms of a fixed signal-to-noise ratio. A similar approach was described in [11, 12, 13] for estimating the probability of speech presence.

Further, the approach by Fisher et al. [1] estimates the noise covariance matrix, which is required in the employed likelihood functions, using a single observation from the input signal. As no averaging is utilized, this estimate is expected to be highly variant. In contrast to [1], we estimate the noise variance using a state-of-the-art noise tracker and explicitly model the statistics of the segment energy, while in [1] the statistics of individual time-samples are modeled. Similar to [1], two models for voiced and unvoiced speech are derived for this feature and compared using a likelihood ratio (LR).

In the following sections, we will first give a mathematical description of the harmonic model and recapitulate the method proposed by Fisher et al. [1] in Section 2. Then, our proposed method is derived in Section 3, followed by a comparison of both approaches in Section 4. Section 5 concludes this paper.

2. SIGNAL MODEL AND PREVIOUS WORK

In this section, we recapitulate the basic structure of the VDA proposed in [1].

For distinguishing voiced and unvoiced segments in noisy conditions, the algorithm exploits the harmonic structure of voiced sounds. In voiced speech, the signal is consequently modeled as noise superimposed by sinusoidal oscillations. In contrast to that, the model for unvoiced speech and silence assumes that only background noise is present. These assumptions are reflected by the hypotheses H_V and H_U, as [1]

$$H_V : \mathbf{x} = \mathbf{A}(\omega_0)\mathbf{b} + \mathbf{d} \tag{1}$$

$$H_U : \mathbf{x} = \mathbf{d}. \tag{2}$$

Here, $\mathbf{x} = [x_1, \ldots, x_N]^T$ and $\mathbf{d}$ represent vectors of N subsequent samples of the input and noise signal, respectively, where $(\cdot)^T$ denotes the vector transpose. The harmonic matrix $\mathbf{A}(\omega_0)$ depends on the fundamental frequency $\omega_0 = 2\pi f_0$. The matrix can be split into

This work was funded by the PhD program “Signals and Cognition”, the DFG Cluster of Excellence EXC 1077/1 “Hearing4all”, and the DFG Research Grant GE2538/2-1.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

$\mathbf{A}(\omega_0) = [\mathbf{A}^c(\omega_0), \mathbf{A}^s(\omega_0)]$ with

$$[\mathbf{A}^c(\omega_0)]_{n,k} = \cos(\omega_0 kn/f_s) \qquad k = 0, \ldots, K \tag{3}$$
$$[\mathbf{A}^s(\omega_0)]_{n,k} = \sin(\omega_0 kn/f_s) \qquad k = 1, \ldots, K \tag{4}$$

and $n = 1, \ldots, N$ [1, 14]. K denotes the number of harmonics, f_s the sampling frequency and the vector $\mathbf{b} = [b_0^c, \ldots, b_K^c, b_1^s, \ldots, b_K^s]^T$ controls the amplitude and phase of the sinusoidal components. $\mathbf{d}$ is assumed to be Gaussian distributed with zero mean and covariance $\mathbf{R_d}$. Fisher et al. [1] additionally assume that this matrix can be decomposed as $\mathbf{R_d} = \mathbf{\Phi} + \sigma_w^2 \mathbf{I}$, where $\mathbf{\Phi}$ is a non-negative definite matrix and σ_w^2 the white noise variance. $\mathbf{I}$ denotes the identity matrix.

Using the two hypotheses, Fisher et al. [1] formulate the LR

$$\Lambda_{\text{Fisher}} = \frac{\max\limits_{\omega_0, \mathbf{b}, \mathbf{R_d}} f(\mathbf{x}|H_V; \omega_0, \mathbf{b}, \mathbf{R_d})}{\max\limits_{\mathbf{R_d}} f(\mathbf{x}|H_U; \mathbf{R_d})} \tag{5}$$

and after deriving the maximum likelihood (ML) estimates for the parameters in both models, the authors obtain the expression

$$\Lambda_{\text{Fisher}} = \sqrt{\frac{\|\mathbf{x}\|^2}{\|\{\mathbf{I} - \mathbf{P_A}(\omega_0)\}\,\mathbf{x}\|^2}}. \tag{6}$$

$\mathbf{P_A}(\omega_0) = \mathbf{A}(\omega_0)\left(\mathbf{A}^T(\omega_0)\mathbf{A}(\omega_0)\right)^{-1}\mathbf{A}^T(\omega_0)$ is a matrix which projects the input vector $\mathbf{x}$ onto the harmonic subspace. For obtaining a decision whether the frame is voiced or unvoiced, the LR is compared to a threshold value. In case it is exceeded, the frame is assumed to be voiced. Otherwise, the frame is classified as unvoiced.

In [1] a constrained ML estimate for $\hat{\mathbf{R}}_\mathbf{d}$ is employed which was previously derived in [15]. Particularly, the covariance matrix is estimated by using a single observation vector $\mathbf{x}\mathbf{x}^T$. No statistical expectation is used here and the zero valued eigenvalues of the singular matrix are set to a lower threshold for which the white noise variance σ_w^2 is chosen. This estimate may describe the noise statistics only rudimentary, since an estimate based on a single data sample suffers from a high variance. It should be noted that the ML estimator for the fundamental frequency and the coefficients for the harmonic tone complex under the voiced hypothesis are the same as in [14]. They are given by

$$\hat{\omega}_0 = \arg\max_{\omega_0} \|\mathbf{P_A}(\omega_0)\mathbf{x}\|^2 \tag{7}$$
$$\hat{\mathbf{b}} = \left(\mathbf{A}^T(\omega_0)\mathbf{A}(\omega_0)\right)^{-1}\mathbf{A}^T(\omega_0)\mathbf{x}. \tag{8}$$

With Bayes' theorem, the voiced probability $p(H_V|\mathbf{x})$ can be obtained from the LR via

$$p(H_V|\mathbf{x}) = \frac{\eta\Lambda}{1 + \eta\Lambda}, \tag{9}$$

where $\eta = p(H_V)/p(H_U)$ is the ratio of the prior probabilities, which are denoted by $p(H_V)$ for the voiced and $p(H_U) = 1 - p(H_V)$ for the unvoiced hypothesis.

With respect to the estimation of the a posteriori probability $p(H_V|\mathbf{x})$, the signal models proposed in [1] exhibit the shortcoming that the resulting likelihood functions for the voiced and unvoiced hypothesis become equal in unvoiced speech. This is because the unvoiced speech signal is not explicitly modeled in the unvoiced state H_U (2). As a consequence, the statistical models for voiced and unvoiced speech cannot be discriminated and the smallest value obtainable for the LR is 1. As a result, also $p(H_V|\mathbf{x})$ is bounded and its lower limit corresponds to the a priori probability $p(H_V)$, which is often chosen to be 50 %. This behavior can also be seen from (6), where the ML optimal estimates of all parameters have been employed in the likelihood functions. If $\mathbf{x}$ does not contain any harmonic components, the denominator in (6) equals approximately $\|\mathbf{x}\|^2$ and thus Λ_{Fisher} is bounded with $\Lambda_{\text{Fisher}} \geq 1$. This may pose a drawback for algorithms which rely on this probability, e. g. a vocoder which mixes the synthesized components for voiced and unvoiced speech based on the a posteriori probability, or speech enhancement algorithms [3].

3. PROPOSED METHOD

In this section we propose a novel method for estimating $p(H_V|\mathbf{x})$. Especially, we will go into detail about the countermeasures taken in order to avoid overlapping likelihood functions and describe an alternative approach for estimating the noise power.

For this approach we suppose that the speech signal $\mathbf{s}$ can be decomposed into a harmonic component $\mathbf{s}_h$ and a non-harmonic component $\mathbf{s}_u$, such that $\mathbf{s} = \mathbf{s}_h + \mathbf{s}_u$. $\mathbf{s}_h$ can be described by the harmonic model and therefore we set $\mathbf{s}_h = \mathbf{A}(\omega_0)\mathbf{b}$. The model parameters ω_0 and $\mathbf{b}$ are determined by the ML estimators in (7) and (8), respectively [1, 14]. The non-harmonic part $\mathbf{s}_u$ is assumed to be a random signal which can be characterized by a Gaussian distribution with zero-mean and variance σ_u^2. Following that, the input signal can be written as

$$\mathbf{x} = \mathbf{s}_h + \mathbf{s}_u + \mathbf{d} = \mathbf{P_A}(\hat{\omega}_0)\mathbf{x} + \mathbf{s}_u + \mathbf{d}. \tag{10}$$

For distinguishing between voiced and unvoiced speech frames we use the statistics of the non-harmonic energy which is given by $\Theta = \|\mathbf{x} - \mathbf{s}_h\|^2$. For the voiced hypothesis we assume that the speech signal is purely harmonic, implying that $\mathbf{s}_u = 0$. On the opposite, we expect no harmonic speech components to be present in unvoiced segments and therefore $\mathbf{s}_h = 0$ is assumed. From this, the following hypotheses can be established:

$$H_V : \Theta = \|\mathbf{x} - \mathbf{s}_h\|^2 = \|\mathbf{d}\|^2 \tag{11}$$
$$H_U : \Theta = \|\mathbf{x}\|^2 = \|\mathbf{d} + \mathbf{s}_u\|^2. \tag{12}$$

Given that the harmonic model perfectly holds, the expected value, $\mathbb{E}(\cdot)$, for the two hypotheses can be computed according to

$$\mathbb{E}(\Theta|H_V) = \mathbb{E}\left(\|\mathbf{d}\|^2\right) = N\sigma_d^2 \tag{13}$$
$$\mathbb{E}(\Theta|H_U) = \mathbb{E}\left(\|\mathbf{d} + \mathbf{s}_u\|^2\right) = N\sigma_d^2(1 + \xi_u). \tag{14}$$

As we assume that the noise and the unvoiced speech signal origin from two independent sources, the two quantities are uncorrelated. $\xi_u = \sigma_u^2/\sigma_d^2$ denotes the signal to noise ratio (SNR) for unvoiced speech and is considered to be fixed. We argue that, similar to [11, 13], this value should represent the SNR which is expected in unvoiced segments. This parameter separates the voiced and unvoiced likelihood functions thus allowing for low a posteriori probabilities $p(H_V|\mathbf{x})$ in unvoiced speech.

If the noise $\mathbf{d}$ is Gaussian distributed with zero mean, Θ is a sum of squared Gaussian distributed random variables and can be modeled by a χ^2 distribution [2], which is a special case of the Gamma distribution. Its mathematical description is

$$f_{\chi^2}(\Theta|\mu, \nu) = \begin{cases} \left(\frac{\nu}{\mu}\right)^{\nu} \frac{\Theta^{\nu-1}}{\Gamma(\nu)} \exp\left(-\frac{\nu}{\mu}\Theta\right) & \text{if } \Theta \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

In this definition $\mu = \mathbb{E}(\Theta)$ is the mean value of the distribution and the parameter ν characterizes the shape of the distribution. The shape depends on the correlation of the samples before being squared and summed. It is given by the relation between the variance and the squared mean

$$\nu = \mathrm{Var}(\Theta)/\mu^2. \tag{16}$$

Using (15) a likelihood ratio can be derived. If the parameter μ is set to the expected values given in (13) and (14), we obtain

$$\begin{aligned}\Lambda =& \frac{f_{\chi^2}(\Theta|\mathbb{E}(\Theta|H_V),\nu)}{f_{\chi^2}(\Theta|\mathbb{E}(\Theta|H_U),\nu)} && (17)\\ =&(1+\xi_u)^\nu \left(\frac{\|\mathbf{x}-\mathbf{P}_{\mathbf{A}}(\hat{\omega}_0)\mathbf{x}\|^2}{\|\mathbf{x}\|^2}\right)^{\nu-1} \times \\ &\exp\left(\frac{\nu}{N\sigma_d^2}\left\{\frac{\|\mathbf{x}\|^2}{1+\xi_u} - \|\mathbf{x}-\mathbf{P}_{\mathbf{A}}(\hat{\omega}_0)\mathbf{x}\|^2\right\}\right). && (18)\end{aligned}$$

In order to evaluate (18) the noise variance σ_d^2 and the shape parameter ν need to be determined. For estimating the noise variance σ_d^2 we employ the speech presence probability (SPP) based estimator proposed in [12]. This approach, however, determines the noise variance for each time-frequency bin in the short-time Fourier domain. In order to obtain an estimate for the variance of a frame in the time domain, we assume that the noise is stationary within a short time period and compute the mean noise power from the spectral estimates. As the employed inverse Fourier transform is normalized by N, we obtain with Parseval's theorem

$$\sigma_d^2 = \frac{1}{N^2}\sum_{k=1}^{N} \sigma_{D,k}^2. \tag{19}$$

$\sigma_{D,k}^2$ denotes the variance of the Fourier coefficients, while σ_d^2 denotes the mean variance in the time domain for the respective segment. If we now employ the variance estimates of the SPP based noise power spectral density (PSD) estimator [12], equation (19) gives an estimate of the noise variance σ_d^2 in the time domain.

The shape parameter ν can be determined by exploiting the relationship between the squared mean and the variance (16) which holds for χ^2 distributed variables. If the expected values in equation (16) are replaced by sample estimates, ν can be trained for different background noises. Under the assumption, that additional unvoiced speech energy changes the shape of the distribution only marginally, the necessary data can be obtained by computing $\|\mathbf{d}\|^2$ for many time instants on different noise signals.

However, for better generalization, we propose to estimate this parameter on-line using recursive filtering, as

$$\mathbb{E}(\Theta) \approx \hat{\mu}[\ell] = (1-\alpha)\hat{\sigma}_d^2[\ell] + \alpha\hat{\mu}[\ell-1] \tag{20}$$

$$\mathbb{E}(\Theta^2) \approx \hat{\beta}[\ell] = (1-\alpha)(\hat{\sigma}_d^2[\ell])^2 + \alpha\hat{\beta}[\ell-1] \tag{21}$$

$$\hat{\nu}[\ell] = \left(\hat{\beta}[\ell] - \hat{\mu}^2[\ell]\right)/\hat{\mu}^2[\ell]. \tag{22}$$

Here, ℓ is the frame index and $\hat{\sigma}_d^2[\ell]$ the noise variance in the time domain estimated using the SPP based approach. Also, here we make the assumption that the shape is not influenced by additional unvoiced speech energy. Again, (9) can be used for estimating the a posteriori probability from the LR in (18).

4. EXPERIMENTAL RESULTS

In this section we compare the proposed approach to the method proposed by Fisher et al. [1]. First, we will describe the audio data and error measures that we use in our experiments. Following that, we will show the different behaviors of both algorithms using results obtained from a single speech utterance. Last, we present the outcome of an experiment which was conducted on a large set of audio files.

For our evaluation we use utterances from the TIMIT corpus [16] which are artificially corrupted by background noises taken from the NOISEX-92 database [17]. We employ pink and babble noise which are mixed at SNRs ranging from 0 to 20 dB in 5 dB steps. Each utterance is extended by a 1.5 s long part of noisy speech in the beginning to allow the on-line estimation of the shape parameter ν (22) to converge sufficiently. This additional part is not considered in the evaluation. Because the TIMIT database contains twice as many sentences by male speakers as sentences by female speakers, we remove one half of the male utterances in order to avoid the results to be biased to one of the two genders. In all our experiments, we use a sampling rate of $f_s = 16$ kHz.

The ground truth labels for the voiced segments are obtained from the pitch annotation provided by [18]. As the annotation is only available for the TIMIT core set, we restrict the experiments to this part of the database. To distinguish unvoiced speech segments from silence periods, we determine the regions of speech activity within the utterances using the clean condition signals. For this, each file is split into 32 ms frames which overlap by 50 %. Afterwards, the power in every frame is computed and divided by the maximum frame power of the respective utterance. If the ratio exceeds -45 dB, the frame is marked as speech active. Speech active frames which have not been marked as voiced within the annotation are consequently labeled as unvoiced.

For assessing the estimation error of the a posteriori probability $p(H_V|\mathbf{x})$, we define the average over-estimation error $\rho_\uparrow$ and under-estimation error $\rho_\downarrow$ as

$$\rho_\uparrow = \frac{1}{|\mathbb{U}|}\sum_{\ell\in\mathbb{U}} \hat{p}(H_V|\mathbf{x}_\ell)\,. \tag{23}$$

$$\rho_\downarrow = \frac{1}{|\mathbb{V}|}\sum_{\ell\in\mathbb{V}} 1-\hat{p}(H_V|\mathbf{x}_\ell)\,. \tag{24}$$

Here, the vector $\mathbf{x}_\ell$ consists of the samples of the ℓth frame. Furthermore, $\mathbb{V}$ is the set of the voiced speech frame indexes, whereas $\mathbb{U}$ incorporates all frame indexes where unvoiced speech is present excluding silence periods. $|\cdot|$ is the cardinality of a set and $\hat{p}(H_V|\mathbf{x}_\ell)$ denotes the estimated posterior. In unvoiced speech, it should be as close as possible to 0 %. Consequently, estimates larger than 0 % contribute to the over-estimation error. Similarly, $\hat{p}(H_V|\mathbf{x}_\ell)$ should be close to 100 % in voiced speech and therefore smaller estimates are understood as under-estimation errors as shown in (24). The introduced error measures are closely related to the false alarm rate (FAR) and missed hit rate (MHR), which are often used for evaluating VDAs. In contrast to the FAR and MHR, here, the hard decision, whether a segment has been detected correctly or not, is replaced by the estimated a posteriori probability.

For our experiments we choose a frame length of 32 ms and a frame shift of 16 ms. The fundamental frequency is determined using (7) by testing all possible candidates on a 1 Hz grid, where the search range is limited to frequencies between 50 and 400 Hz.

We use the utterances from TIMIT's core training set to optimize the number of harmonics K and the unvoiced SNR ξ_u. With the removal of one half of the male utterances, 24 sentences are available for each gender, each noise type and each SNR. Depending on the background noise, we select the parameters which yield the smallest error rates with respect to (23) and (24). For obtaining the optimal settings, we employ a brute force testing procedure based on the training data. As a result of this procedure, the number of harmonics K is set to 5 for both algorithms and for the unvoiced SNR $\xi_u = -5$ dB is used in pink noise and 0 dB in babble noise. In all our evaluations,

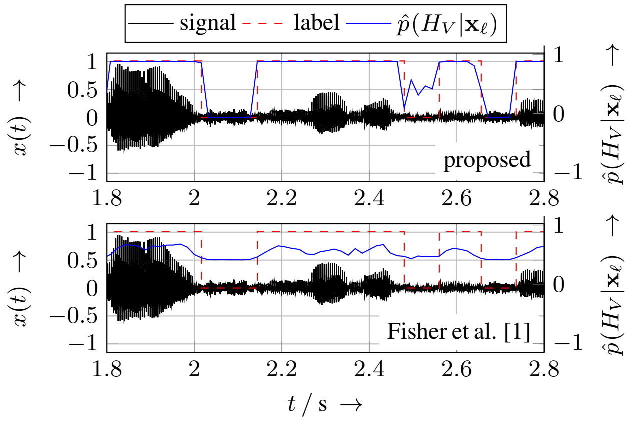


Fig. 1: A posteriori probability estimated using the proposed method with recursively estimated shape parameter (top) and the approach proposed by Fisher et al. [1] (bottom) for a speech signal in pink noise at 15 dB SNR, where the label indicates voiced (1) and unvoiced (0) segments (as provided in [18]).

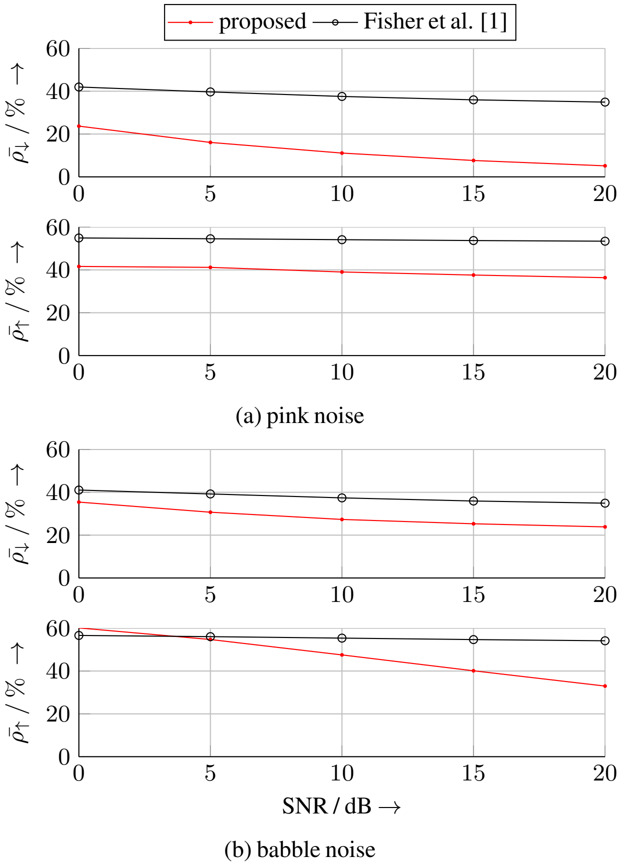
we estimate the shape parameter ν during processing using (20) - (22). For this, we employ the smoothing constant $\alpha = 0.98$.

Figure 1 depicts the estimated a posteriori probability $\hat{p}(H_V|\mathbf{x}_\ell)$ for the proposed method and the approach given by Fisher et al. [1]. The a priori probability for voiced and unvoiced speech is assumed to be equal and the employed speech signal was degraded by pink noise at 15 dB SNR. The figure shows that the probabilities obtained from the approach by Fisher et al. [1] do not fall below 50 %. This behavior is expected because, as discussed above, the likelihood functions for both hypotheses cannot be distinguished in unvoiced segments. Furthermore, it can be observed that the occurrence of a voiced speech segment gives only a subtle rise with respect to the estimated voicing probability. In contrast to that, the a posteriori probabilities estimated by the proposed approach reveal values close to 100 % during voiced segments and values near 0 % in unvoiced parts. In uncertain frames, e. g. in Figure 1 around 2.5 s, the posterior probability is between 0 and 1.

In the remainder of this section, we will focus on the evaluation of the average over-estimation and under-estimation error, which were computed on the TIMIT core test set. For this, we compute (23) and (24) for each utterance. Afterwards, these results are averaged over all sentences for each SNR and for each noise type. The a priori probabilities for the voiced and unvoiced hypotheses are again assumed to be equal, i. e. $p(H_V) = p(H_U)$. Here, there are 64 sentences available for each gender, noise type and SNR. Figure 2 shows the outcome for pink and babble noise for both algorithms.

From the figure it is visible that the under-estimation error for the proposed method is considerably lower than for the approach by Fisher et al. [1] which agrees with the findings from Figure 1. The proposed method assigns probabilities close to 1 in voiced segments, whereas only small increases from 0.5 in the estimated posterior probability are observed for the algorithm by Fisher et al. [1].

The proposed method can also achieve lower over-estimation errors as shown in Figure 2. However, the over-estimation error obtained for the proposed method is considerably larger than the under-estimation error revealing the algorithm's general tendency towards overestimating $p(H_V|\mathbf{x})$. In babble noise for example, the over-estimation error is slightly worse at 0 dB SNR compared to the approach by Fisher et al. [1]. In such noisy conditions, the proposed algorithm is not able to assign low probabilities for unvoiced segments properly, which leads to a behavior similar to the approach by Fisher et al. [1].



(a) pink noise

(b) babble noise

Fig. 2: Mean of the over-estimation $\bar{\rho}_\uparrow$ and under-estimation error $\bar{\rho}_\downarrow$ for the proposed approach and the algorithm by Fisher et al. [1] averaged over all utterances in dependence on the SNR in pink noise and babble noise. The averaging is indicated by $\bar{\cdot}$.

In higher SNRs conditions, these problems do not occur and low probabilities are assigned to unvoiced segments. This is reflected by lower over-estimation errors which are also smaller compared to the approach by Fisher et al. [1]. In contrast to the proposed algorithm, this method yields an over-estimation error slightly above 50 % over all SNR conditions. This reflects again the behavior that the estimated posterior probabilities are restricted to a lower bound which is given by the a priori probability of voiced speech (e. g. 0.5). According to that, the proposed method can outperform the algorithm by Fisher et al. [1]. Further, these results indicate that the recursive smoothing in (20) - (22) obtains reasonable estimates for the shape parameter ν and can deal with unknown noise types.

5. CONCLUSIONS

In this paper, we presented a new method for estimating the a posteriori probability of a frame being voiced for which a harmonic model is used. In comparison to the approach by Fisher et al. [1] we apply two major changes. While Fisher et al. [1] estimate the noise covariance matrix using a single observation vector, we employ a state-of-the art noise PSD tracker. This is possible, because we use the segment energy in our model for distinguishing voiced and unvoiced sounds. Additionally, we introduce a fixed SNR for unvoiced speech, which allows the proposed method to correctly assign low probabilities during unvoiced speech. In the mean, these modifications lead to more accurate estimations of the a posteriori probability compared to the method proposed by Fisher et al. [1].

6. REFERENCES

[1] Etan Fisher, Joseph Tabrikian, and Shlomo Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 502 – 510, 2006.

[2] Peter Vary and Rainer Martin, *Digital speech transmission: Enhancement, Coding and Error Concealment*, Wiley & Sons, Chichester, West Sussex, UK, 2006.

[3] Martin Krawczyk, Robert Rehr, and Timo Gerkmann, "Phase-sensitive real-time capable speech enhancement under voiced-unvoiced uncertainty," in *21st European Signal Processing Conference 2013 (EUSIPCO 2013)*, Marrakech, Morocco, 2013.

[4] Bishnu S. Atal and Lawrence R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 3, pp. 201–212, 1976.

[5] Huiqun Deng and D. O'Shaughnessy, "Voiced-unvoiced-silence speech sound classification based on unsupervised learning," in *IEEE International Conference on Multimedia and Expo, 2007*, July 2007, pp. 176–179.

[6] A. Michael Noll, "Cepstrum pitch determination," *The Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1585–1591, 1968.

[7] Sassan Ahmadi and Andreas S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, May 1999.

[8] Sira Gonzalez and Mike Brookes, "PEFAC - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

[9] Robert J. McAulay and Thomas F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *International Conference on Acoustics, Speech, and Signal Processing, 1990*, Albuquerque, New Mexico, USA, 1990, pp. 249–252 vol.1.

[10] Robert J. McAulay and Thomas F. Quatieri, "Speech analysis/Synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[11] Timo Gerkmann, Colin Breithaupt, and Rainer Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 910–919, 2008.

[12] Timo Gerkmann and Richard. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011*, New Paltz, New York, USA, 2011, pp. 145–148.

[13] Richard C. Hendriks, Timo Gerkmann, and Jesper Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, vol. 9 of *Synthesis Lectures on Speech and Audio Processing*, 2013.

[14] Joseph Tabrikian, Shlomo Dubnov, and Yulya Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 76–87, 2004.

[15] Kerem Harmancı, Joseph Tabrikian, and Jeffrey L. Krolik, "Relationships between adaptive minimum variance beamforming and optimal source localization," *IEEE Transactions on Signal Processing*, vol. 48, no. 1, pp. 1–12, 2000.

[16] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue, "TIMIT acoustic-phonetic continuous speech corpus," Tech. Rep., Linguistic Data Consortium, Philadelphia, 1993.

[17] Herman J. M. Steeneken and Frank W. M. Geurtsen, "Description of the RSG.10 noise database," Technical Report IZF 1988-3, TNO Institute for perception, 1988.

[18] Sira Gonzalez and Mike Brookes, "Pitch of the core TIMIT database set," `http://www.ee.ic.ac.uk/hp/staff/dmb/data/TIMITfxv.zip`, 2011.