PAPER

Continuous Noise Masking Based Vocoder for Statistical Parametric Speech Synthesis

Mohammed Salah AL-RADHI[†a)], ***Student Member***, **Tamás Gábor CSAPÓ**[†,††b)], ***and* Géza NÉMETH**[†c)], ***Nonmembers***

SUMMARY In this article, we propose a method called "continuous noise masking (cNM)" that allows eliminating residual buzziness in a continuous vocoder, i.e. of which all parameters are continuous and offers a simple and flexible speech analysis and synthesis system. Traditional parametric vocoders generally show a perceptible deterioration in the quality of the synthesized speech due to different processing algorithms. Furthermore, an inaccurate noise resynthesis (e.g. in breathiness or hoarseness) is also considered to be one of the main underlying causes of performance degradation, leading to noisy transients and temporal discontinuity in the synthesized speech. To overcome these issues, a new cNM is developed based on the phase distortion deviation in order to reduce the perceptual effect of the residual noise, allowing a proper reconstruction of noise characteristics, and model better the creaky voice segments that may happen in natural speech. To this end, the cNM is designed to keep only voice components under a condition of the cNM threshold while discarding others. We evaluate the proposed approach and compare with state-of-the-art vocoders using objective and subjective listening tests. Experimental results show that the proposed method can reduce the effect of residual noise and can reach the quality of other sophisticated approaches like STRAIGHT and log domain pulse model (PML).

key words: *noise masking, continuous vocoder, speech synthesis, phase distortion, kernel density functions*

1. Introduction

With the fast growth of computer technology to become more functional and prevalent as time passes, a wide range of the speech processing area (such as speech synthesis, speech recognition, dialogue management, etc.) is becoming a core function for establishing a human-computer communication interface. Speech synthesis is the ability to build natural-sounding synthetic voices from text [1]. During the last decades, there are several speech models, which allow the machine to produce spoken responses. However, these synthesis systems are still far from the goal of reaching completely human-sounding speech.

Statistical Parametric Speech Synthesis (SPSS) systems based on a parameterization of the speech waveform (vocoding) have achieved popularity over the last few years [2]. Due to the statistical modelling process, several vocoders have been successfully applied to various kinds of applications such as text-to-speech (TTS) synthesis [3] and voice conversion [4]. However, its biggest drawback is the quality of the synthesized speech [5]. Fortunately, there are new generative models of audio data that operate directly at the waveform level without using a vocoder, such as WaveNet [6]. Even though WaveNet yields state-of-the-art performance and gives a good sounding speech in a variety of voices, it requires a large quantity of data and computation power which makes it difficult to implement and train. Therefore, vocoder-based SPSS still provides a quick and flexible solution that can capture high quality synthesized speech.

A typical vocoder-based SPSS decomposes the speech waveform into a spectral envelope and excitation parameters (e.g. fundamental frequency) to be modeled and modified in a unified framework. Many different vocoders have been proposed over recent years. In the context of high-quality speech synthesis, STRAIGHT [7] and WORLD [8] are the state-of-the-art vocoders to synthesize the voice that sounds as natural as the input voice. However, their high computational complexity and variable parameters are still considered challenging issues, which present some speech quality degradation in the TTS and other speech applications [9]. In our recent work in SPSS [10], we proposed a vocoder using continuous fundamental frequency (contF0) in combination with maximum voiced frequency (MVF), which was successfully used with a deep neural network based TTS [11]. The advantage of a continuous vocoder is that these vocoder parameters are simpler to model than conventional vocoders with discontinuous F0. Similarly to other vocoders (e.g. a lack of proper noise modeling in STRAIGHT [12]), the noise component in continuous vocoder is still not accurately modelled that limits the overall perceived quality. To mitigate the problem above, we propose here a continuous noise masking (cNM) approach with the aim of improving the naturalness of synthetic speech. This method has a twofold advantage: a) it allows to mask out most of the noise residuals; and b) it attempts to reproduce the voiced and unvoiced (V/UV) regions more precisely, that is, resembles natural sound signal. Thus, proper reconstruction of noise in voiced segments (like in breathiness parts) is necessary for the synthetic speech to achieve a quality closer to that of the natural sound.

Noise masking has been widely used in earlier

Manuscript received June 14, 2019.
Manuscript revised October 23, 2019.
Manuscript publicized February 10, 2020.

[†] The authors are with Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, 1117 Budapest, Hungary.

[††] The author is with MTA-ELTE Lendület Lingual Articulation Research Group, 1088 Budapest, Hungary.

a) E-mail: malradhi@tmit.bme.hu
b) E-mail: csapot@tmit.bme.hu
c) E-mail: nemeth@tmit.bme.hu

DOI: 10.1587/transinf.2019EDP7167

studies. One simple method is presented in [13] as a small amount of artificial noise is added to the clean speech to improve the noise immunity of the model and reach the desired signal-to-noise ratio (SNR). Another method with similar goals is capable of lowering the statistical mismatch of acoustic features in the training and testing conditions [14]. Moreover, a good degree of noise robustness in both filter bank and Mel-frequency cepstral domains can be found in [15]. Recently, a binary noise mask has been proposed for improving both speech intelligibility based on noise distortion constraints [16], and parametric speech synthesis based on thresholding the Phase Distortion Deviation (PDD) [17]. However, forcing the PDD values below thresholding to zero might lack a minimum of randomness in the voiced segments [12], [18]. Therefore, by extending the idea examined in [12], we propose a cNM in this article to avoid any residual buzziness, improve creakiness, and ensure the proper randomization of the noise segments in our continuous vocoder.

Objective and subjective evaluation of the improved version of the continuous vocoder is performed. We show that the proposed cNM is superior to our earlier residual-based vocoder. Specifically, it resynthesizes voiced segments more accurately, in which higher improvement in the segmental SNR can be obtained. The rest of the paper is structured as follows. In Sect. 2, we present the theory behind the continuous noise masking. Section 3 evaluates the proposed method, which validates its effectiveness. Finally, we summarize this paper in Sect. 4 and suggest avenues for future research.

2. Proposed Methodology

This section will show our main contributions of improving the latest version of the continuous vocoder [10] by firstly giving the principle idea of how parameterizes and reconstructs speech signals in our vocoder, and secondly introducing the novelty of a continuous noise masking.

2.1 Baseline Vocoder Description

For a better understanding of what is next, the analysis and synthesis phases of continuous vocoder are shown in Fig. 1. During the analysis phase, continuous fundamental frequency (contF0) is calculated on the input waveforms using a simple continuous pitch tracker [19]. In areas of creaky voice, and in the event of unvoiced sounds or silences, this pitch tracker interpolates F0 based on a linear dynamic system and Kalman smoothing. Another excitation parameter is the maximum voiced frequency (MVF) which exploits both amplitude and phase spectra that integrated into a maximum likelihood criterion to derive the MVF decisions [20]. Additionally, 24-order Mel-Generalized Cepstral analysis (MGC) [21] is performed on the speech signal with alpha = 0.58 and gamma = 0. The frameshift is 5ms and the sampling frequency is 16kHz. The results are the contF0, MVF, and MGC parameter streams. The Glottal Closure Instant (GCI) algorithm [22] is used to find the glottal period boundaries of individual cycles in the voiced parts of the inverse filtered residual signal. From these pitch cycles, a principal component analysis (PCA) residual is finally built which will be used in the synthesis phase.

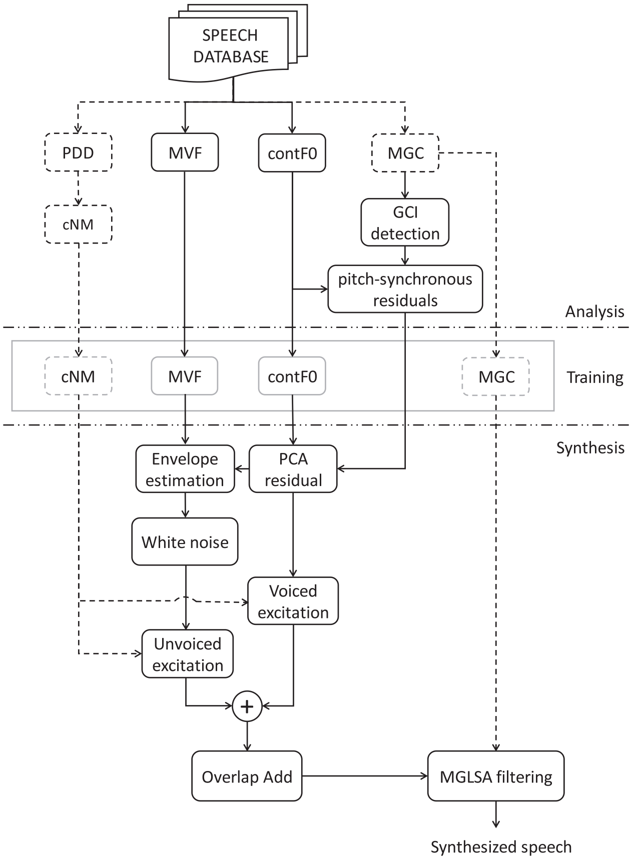


Fig. 1 Schematic diagram of the developed continuous vocoder. Additions and refines are marked with dashed lines.

During the synthesis phase, voiced excitation is made of PCA residuals overlap-added pitch synchronously. This voiced excitation is lowpass filtered frame by frame at the frequency given by the MVF parameter. In the frequencies higher than the actual value of MVF, white noise is applied. Voiced and unvoiced excitation is combined together, and the MGLSA (Mel-Generalized Log Spectrum Approximation) filter is used to synthesize speech [23].

The continuous vocoder has the obvious advantage of avoiding voicing decision per frame that may be considered to reduce the perceptual degradation caused by voicing decision errors. Moreover, it uses only two one-dimensional parameters for modeling the excitation, which is computationally feasible in the deep neural network based text-to-speech [11], [24]. However, and similar to STRAIGHT, the noise component in the baseline continuous vocoder is still not accurately modelled that limits the overall perceived quality. The next section, therefore, explores the way to tackle this issue by proposing a method known as a noise masking.

2.2 Continuous Noise Masking

The aim of this work is to reduce the overall buzziness of the voice that is visible in our baseline continuous vocoder [10]. Recent progress in the synthesized speech showed that the Phase Distortion Deviation (PDD) of the signal carries all of the crucial information relevant to the glottal pulses shape [17]. Moreover, noise masking is a fundamental technique to improve the performance of the speech synthesizer by reducing the number of noise artifacts in the time-frequency domain. In [12], a binary noise masking (bNM) in the time-frequency space is used based on a simple measure of harmonicity. However, bNM might lack a minimum of randomness in the voiced segments because of forcing values below the threshold to zero [12], [18]. Therefore, considering that both PDD and bNM help in decreasing the variability of the speech signal, we propose a new masking approach called continuous noise masking (cNM) that changes from 0 to 1 (or 1 to 0) rather than a binary 0 or 1 as in the bNM, and hence preserves the quality of the voiced segments.

In order to compute the cNM, we should first compute the PDD. Originally, PDD can be calculated based on early Fisher's standard-deviation [25]. However, [17] shows two issues related to variance $\sigma(f)$ and source shape in voiced segments. By avoiding these limitations, PDD can be estimated in this experiment at 5 ms frameshift by

$$PDD = \sigma_i(f) = \sqrt{-2\, log \left| \frac{1}{N} \sum_{n \in C} e^{j(PD_n(f) - \mu_n(f))} \right|} \tag{1}$$

$$\mu_i(f) = \angle \left(\frac{1}{N} \sum_{n \in C} e^{jPD_n(f)} \right) \tag{2}$$

where $C = \left\{ i - \frac{N-1}{2}, \ldots, i + \frac{N-1}{2} \right\}$, N is the total number of frames, $PD(f)$ is the phase difference between two consecutive frequencies f components, i is the frame index, and we denote the phase by $\angle$. As we wanted to quantify the noisiness in the higher frequency bands only, we zeroed out the PDD values below the MVF contour.

Unlike in bNM [12] which was just a thresholded version of PDD, cNM can be estimated here as

$$cNM = 1 - P\acute{D}D(f) \tag{3}$$

where $P\acute{D}D$ is a normalized PDD value using nearest-neighbor resampling method. Then, to model the speech signal in continuous vocoder, the following formulas are applied in the synthesis phase $s(t)$ as shown in Fig. 1:

$$s(t) = \sum_{n=1}^{N} v_n(t) + u_n(t) \tag{4}$$

where $v(t)$ and $u(t)$ are the voiced and unvoiced speech components at frame n, respectively. Thus, for $\forall t$

$$v_n(t) = \begin{cases} v_n(t), & cNM \leq threshold \\ 0, & cNM > threshold \end{cases} \tag{5}$$

$$u_n(t) = u_n(t) * cNM(t) \tag{6}$$

The masking algorithm developed here is to carry out the masking in the voiced and unvoiced segments of the continuous vocoder. To better understanding of how to approach the above conditions, the suggested model shall satisfy the properties: If the value of the cNM estimate for the voiced frame is greater than the threshold, then this value is replaced (masked) in order to reduce the perceptual effect of the residual noise as may appear in the voiced parts of the cNM (lower values), whereas Eq. (6) controls the unvoiced frame based on the unvoiced part of the cNM (higher values). This means that cNM can save parts of speech component in the weak-voiced and unvoiced segments by using a smaller value instead of 0 or 1 caused by the bNM estimation.

Accordingly, cNM improves the synthesis robustness to noise generated in creaky voice segments and closely resembles natural background noise (such as breathy voice). In informal listening tests, we experimented with several continuous values (from 0 to 1), and selected 0.77 as the one producing the best results for indication of presence/absence of voicing in respective voiced/unvoiced frames. This threshold is supported by validated the experiment in Sect. 3 (Fig. 4) showing that the probability kernel density function of the proposed model (blue line) starts to match the natural one (black dash line) at PDD 0.77, which then is confirmed as a confidence threshold in this article to avoid any other erroneous estimates. Nevertheless, the results are not to be very sensitive to this threshold as it is more like a clipping needed to account for a low and high level estimation issue in the voiced and unvoiced frames. Future works might focus on more elaborated strategies, i.e. a user-definable threshold in graded form reflecting different levels of voice and noise strength. Moreover, the masking threshold is not data dependent in our article as it generalizes to other datasets as well. However, further validation of this study is required on other languages.

An example of cNM estimation on a female speech sample and masking threshold compared with the MVF contour is shown in Fig. 2. It can be seen that the cNM also follows the actual voiced/unvoiced regions of the MVF. In other words, if the segment is voiced, then the cNM must be lower value to give indication to the synthesis process that this region is voiced and should discard any other noise artifacts depends on the threshold. On the contrary, if the segment is unvoiced, then the cNM must be higher value to give indication to the synthesis process that this region is unvoiced and should mask any other higher harmonics frequencies depends also in the threshold. Consequently, it possible for this method to reduce the effect of residual noise, and thus yielding to save parts of speech components.

2.3 Mel-Generalized Cepstral Algorithm

In our recent studies [10], a simple spectral model represented by 24–order MGC was used [21]. Although several

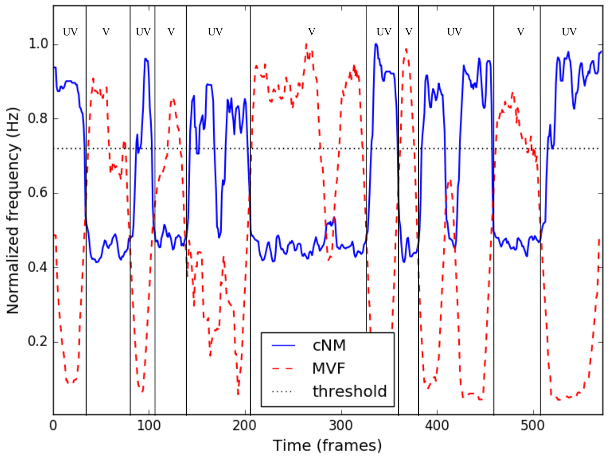


Fig. 2 Illustration of the performance of the continuous noise mask (blue line) plotted across the maximum voiced frequency (red dashed line), where threshold = 0.77 (black dotted line) is obtained after informal listening tests; UV and V are the unvoiced and voiced segments, respectively. English sentence: "I was not to cry out in the face of fear." from a female speaker.

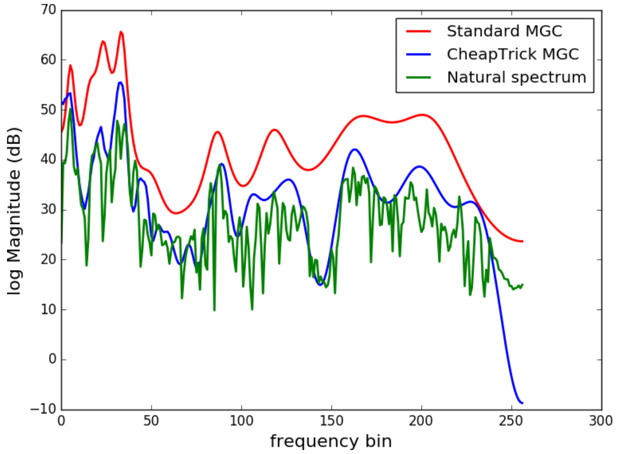


Fig. 3 Example of the signal spectrum of a voiced segment (green) with the spectral shape (spectral envelope) estimates obtained with standard MGC (red) and CheapTrick (blue).

vocoders based on this simple algorithm have been developed, they are not able to synthesize natural sound. The main problem is that it is affected by time-varying components and it is difficult to remove them. Therefore, more advanced spectral estimation methods might increase the quality of synthesized speech.

In [26], an accurate and temporally stable spectral envelope estimation called CheapTrick was proposed. CheapTrick consists of three steps: F0-adaptive Hanning window, smoothing of the power spectrum, and spectral recovery in the quefrency domain. In a modified version of the continuous vocoder, Cheaptrick algorithm using the 60-order MGC representation with $\alpha = 0.58$ (Fs = 16 kHz) will be used to achieve high-quality speech spectral estimation. A comparison of the spectral envelope between standard MGC and the CheapTrick is shown in Fig. 3. Accordingly, it is clear now to see how a continuous vocoder will behave after adaptation to a more accurate spectral envelope technique than the previous MGC system.

3. Experimental Evaluation and Discussion

3.1 Datasets

We used a CMU-ARCTIC database [27] to evaluate the sound quality of the proposed algorithm. The parallel speech data of four speakers are chosen as our corpus, denoted BDL (American English, male), JMK (Canadian English, male), SLT (American English, female), and CLB (US English, female). Each one produced one hour of speech data segmented into 1132 sentences, restricting their length from 5 to 15 words per sentence (a total of 10045 words with 39153 phones). Moreover, CMU-ARCTIC are phonetically-balanced utterances with 100% phonemes, 79.6% diphones, and 13.7% triphones. As sample frequency 16kHz and 16-bit samples are used, and acoustic features were extracted with a 5ms frameshift. In the vocoding experiments, 100 sentences (25 sentences from each speaker) were chosen randomly to be analyzed and synthesized with the baseline [10], STRAIGHT [7], log domain pulse model (PML) [12], and proposed vocoders. In order to reach our points and to prove the effectiveness of the proposed approach, objective and subjective evaluations were carried out in the next subsections.

3.2 Objective Evaluation

Finding a meaningful objective metric is always a challenge in evaluating the performance of speech quality, similarity, and intelligibility. In fact, one metric might be possibly suitable for a few systems but not convenient for all. The reason for that may be returned to some factors which are influenced by the speed, complexity, or accuracy of the speech models. Speaker types and environmental conditions should also be taken into account when choosing these metrics. Therefore, four objective speech quality measures are considered to evaluate the quality of the proposed model. Frequency-weighted segmental signal-to-noise ratio (fwSNRseg) [28] was firstly calculated, defined as

$$fwSNR_{seg} = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\sum_{i=1}^{K} W_{i,j} \cdot log\, \frac{X_{i,j}^2}{X_{i,j}^2 - Y_{i.j}^2}}{\sum_{i=1}^{K} W_{i,j}}\right) \tag{7}$$

where $X_{i,j}^2$, $Y_{i.j}^2$ are critical-band magnitude spectra in the j^{th} frequency band of the target and converted frame signals respectively, K is the number of bands, and W is a weight vector. Secondly, coherence and speech intelligibility index (SII) [29] was employed to evaluate the noise and distortion of the synthetic speech. The coherence SII (CSII) measure was chosen here because it has been shown in [30] to be one

Table 1 Average scores performance based on re-synthesized speech for male and female speakers. The bold font shows the best performance of each column.

Metric	Speaker	Models			
		Baseline	PML	STRAIGHT	Proposed
fwSNRseg	JMK	6.083	9.959	**14.436**	11.661
	BDL	6.449	13.578	**16.371**	12.298
	CLB	7.559	13.752	**16.583**	9.789
	SLT	6.771	13.538	**15.742**	10.938
coherence SII	JMK	0.048	0.208	0.252	**0.271**
	BDL	0.044	0.191	0.244	**0.248**
	CLB	0.043	0.199	**0.226**	0.204
	SLT	0.065	0.236	0.252	**0.263**

of the best predictors for speech intelligibility in fluctuating noise conditions. In this work, the CSII is obtained for each frame m as:

$$CSII_j(m) = 10\log_{10}\frac{\sum_{k=0}^{I-1}\hat{X}(m,k)\cdot W_j(k)}{\sum_{k=0}^{I-1}\hat{S}(m,k)\cdot W_j(k)} \tag{8}$$

where $W_j(k)$ is the filter windows function, k is the FFT bin index, $\hat{X}(m,k)$ and $\hat{S}(m,k)$ are estimations of the natural and synthesized speech power spectra, respectively. These are obtained as

$$\hat{X}(m,k) = |\gamma(k)|^2\cdot|S(mT,k)|^2 \tag{9}$$
$$\hat{S}(m,k) = (1-|\gamma(k)|^2)\cdot|S(mT,k)|^2 \tag{10}$$

where $S(mT,k)$ is the short-time Fourier transform of the synthesized speech, T is the frameshift, and the magnitude squared coherence γ of the cross-spectral density S_{xs} between natural speech $x(n)$ and synthesized speech $s(n)$, both having spectral densities S_{xx} and $S_{ss}(k)$ respectively, is given by

$$|\gamma(k)|^2 = \frac{|S_{xs}|^2}{S_{xx}(k)S_{ss}(k)}, \qquad 0\le|\gamma(k)|^2\le 1 \tag{11}$$

Additionally, the density estimate using a kernel smoothing method [31], [32] was calculated to show how the reconstruction of the noise component in the state-of-the-art vocoders behaved in comparison to the proposed model. The probability kernel density function is given by

$$\hat{f}_h(s) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{s-y_i}{h}\right) \tag{12}$$

where s is the synthesized speech signal, $\{y_i\}_{i=1}^{n}$ are finite random samples drawn from some distribution with an unknown density, $K(\cdot)$ is the kernel function, and $h > 0$ is a smoothing parameter to adjust the width of the kernel. A more detailed case-by-case analysis by fwSNRseg and CSII are shown in Table 1. The results were averaged over 25 synthesized test utterances for each speaker. A calculation is done frame-by-frame, and the best value in each column of Table 1 is boldfaced.

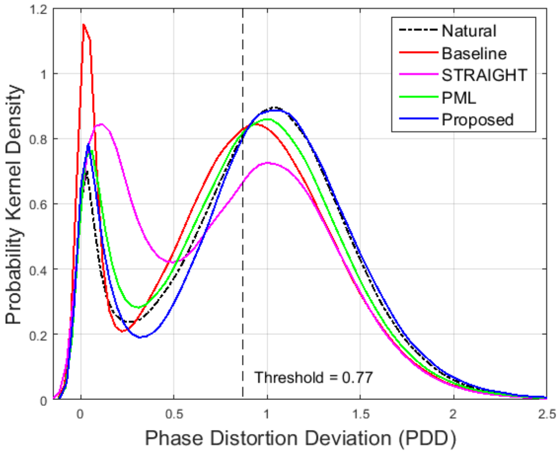


Fig. 4 Estimation of the probability kernel density functions of PDDs using 4 vocoders compared with the PDD measure on the natural speech signal. The threshold = 0.77 is shown in the vertical dashed line.

First, it could be observed that our proposed method significantly outperforms the baseline vocoder in both metrics. In particular, it can be seen from the fwSNRseg measure that the proposed vocoder is also better than PML in the JMK speaker. On the contrary, STRAIGHT vocoder still gives better metric results than other systems. Second, for both male and SLT female speakers, the coherence SII values in Table 1 indicate that the proposed system obviously outperforms all systems. In a sense, there is a tendency to increased CSII when considering continuous noise masking in the proposed method. It is interesting to emphasize that the baseline does not at all meet the performance of the other vocoders in all speakers. In other words, the results reported in Table 1, strongly support the use of proposed vocoder than others in terms of coherence SII measure. We can conclude that the approach reported in this work is beneficial and can substantially reduce any residual buzziness. But it should be pointed out that there is no guarantee that better objective measures yield a better model as synthetic speech quality is an inherently perceptual study.

Probability kernel density function of PDD values for all systems are also estimated and shown in Fig. 4 compared

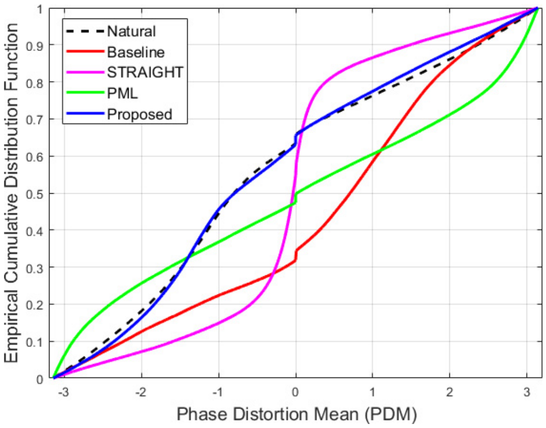


Fig. 5 Empirical cumulative distribution function of PDMs using 4 vocoders compared with the PDM measure on the natural speech signal.

to the PDD measure on the natural speech signals. It can be shown that the proposed vocoder based cNM start to match the natural PDD values at a threshold of 0.77, whereas other systems (like STRAIGHT) presents more deviation from the natural one. This indicated that the proposed cNM method gives a better synthesis of the noise in voiced and unvoiced segments than, for example, the bNM in PML.

Finally, the empirical cumulative distribution function [33] of phase distortion mean values are calculated and displayed in Fig. 5 to see whether these systems can be normally distributed and how far they are from the natural signal. The empirical cumulative distribution function $F_n(PDM)$ defined as

$$F_n(x) = \frac{\#\{X_i : X_i \leq x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}(X_i) \tag{13}$$

where X_i is the PDM variables with density function $f(x)$ and distribution function $F(x)$, $\#A$ symbolizes the number of elements in the set A $(X_i \leq x)$, n is the number of experimental observations, I is the indicator of event A given as

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} \tag{14}$$

It can be noticed that the higher mode of the distribution (positive x-axis in Fig. 5) corresponding to STRAIGHT's PDMs is clearly higher than that of the original signal, while the PML's PDMs is lower. This also demonstrates why the synthesized speech for them ranked lower in the perception test (see Sect. 3.2). On the contrary, the higher mode of the distribution corresponding to the proposed configuration is better synthesized performance with almost matching the natural speech signal. The performance of STRAIGHT and the baseline vocoders appear considerably worse than PML. Focusing on the lower mode of the distribution (negative x-axis in Fig. 5), PML's PDMs gives the second better synthesized performance behind the proposed model. This result is probably explained by the fact that cNM can substantially reduce any residual buzziness.

```
1.  contF0 = Continuous(X, Fs)
2.  MGC = CheapTrick(X, Fs, F0)
3.  MVF = MaxVoiceFreq(X, F0)
4.  cNM = Proposed(PDD)
5.  Y = Synthesis(contF0, MVF, MGC, cNM)
```

Fig. 6 Steps of the continuous vocoder. X represents the input waveform, Fs represents the sampling frequency, and Y represents the synthesized speech.

In general, the experimental results confirm the effectiveness of the proposed vocoder in terms of speech naturalness is comparable, or even better, to the STRAIGHT and PML vocoders. As a summary, the analysis and synthesis steps for the latest version of the continuous vocoder are shown in Fig. 6.

3.3 Subjective Evaluation

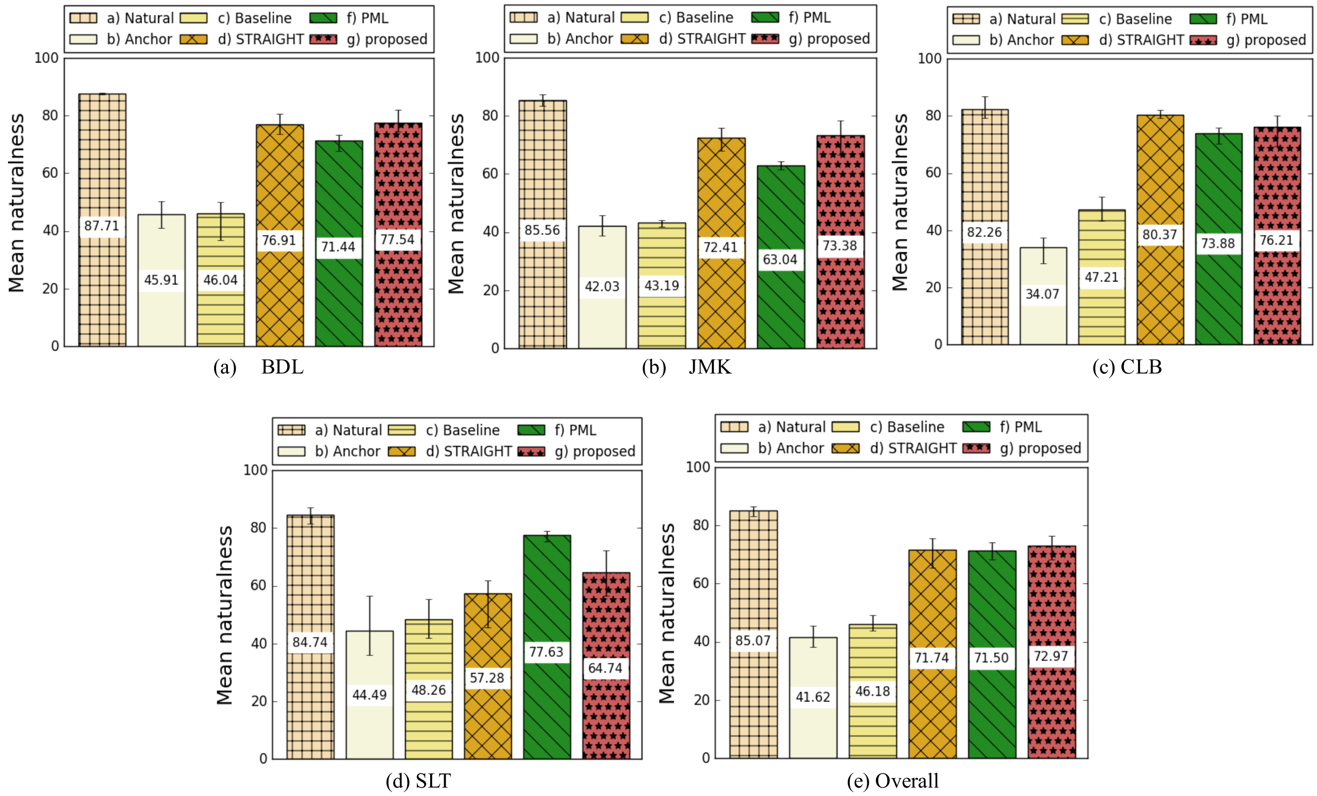
In order to evaluate the perceptual quality of the proposed systems, we conducted a web-based MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening test [34]. We compared natural sentences with the synthesized sentences from the baseline, proposed, STRAIGHT, PML, and an anchor system. The anchor type was the re-synthesis of the sentences with a standard MGLSA vocoder using pulse-noise excitation [23] implemented in speech signal processing toolkit (SPTK)[†]. In the test, the listeners had to rate the naturalness of each stimulus relative to the reference (which was the natural sentence), from 0 (highly unnatural) to 100 (highly natural). The utterances were presented in a randomized order. The listening test samples can be found online[††]. 18 participants (9 males, 9 females) with a mean age of 29 years were asked to conduct the online listening test.

We evaluated 16 sentences (4 from each speaker). Altogether, 96 utterances were included in the test (4 speaker x 6 types x 4 sentences). On average, the test took 12 minutes to fill. The MUSHRA scores for all the systems are shown in Fig. 7, showing both speaker by speaker and overall results.

According to the results, the proposed vocoder clearly outperformed the baseline system (Mann-Whitney-Wilcoxon ranksum test, $p < 0.05$). Particularly, one can see that in the case of both male speakers (BDL and JMK) the proposed method is significantly better than the PML and STRAIGHT vocoders. In terms of the female speakers (Fig. 7 (c), (d)), we can see that the proposed vocoder is ranked as the second best choice. In other words, the vocoder based cNM is superior to the method based bNM in PML and the method based voice decision in STRAIGHT vocoder in case of CLB and SLT speakers, respectively. This unexpected difference (specially in Fig. 7 (d)) probably might be due to one of two concerns. First, SLT

[†] http://sp-tk.sourceforge.net/

[††] http://smartlab.tmit.bme.hu/cNM2019



(a) BDL (b) JMK (c) CLB

(d) SLT (e) Overall

Fig. 7 Results of the subjective evaluation for the naturalness question. A higher value means larger naturalness. Error bars show the bootstrapped 95% confidence intervals.

under-articulates, speaks with a low vocal effort, and exhibit a pressed voice quality [35]. Alternatively, the female SLT speaker has a rather modal phonation with a bit of nasality, which is affected the evaluation scores. Second, the voiced/unvoiced decision was also left up to the Maximum Voiced Frequency parameter in our study, whereas other systems have separate complex parameters to model this (e.g. aperiodicity parameter in STRAIGHT). Therefore, some possibly inaccurate decisions might be also occurred (especially in unvoiced regions). Listeners seem to prefer the female voices of PML and the male voices of the proposed model. But our system is simpler, i.e. uses less parameters compared to STRAIGHT and PML vocoders.

Based on the overall results, we can conclude that among the techniques investigated in the study of noise reconstruction, cNM performs well in continuous vocoder when compared with other approaches (Fig. 7 (e)). When taking these overall results, the difference between STRAIGHT, PML and the proposed system is not statistically significant (Mann-Whitney-Wilcoxon ranksum test, $p < 0.05$), meaning that our methods reached the quality of other state-of-the-art vocoders. This positive result was confirmed by a coherence SII measure in the statistical aspects of the objective's experimental test.

4. Conclusions

This work developed an encouraging alternative method to reconstruct the noisiness of the speech signal in a continuous vocoder. We have described an implementation of how to generate such a continuous noise masking to avoid any residual buzziness. It was also shown in a subjective listening test that the continuous vocoder allows better ability to synthesize the speech compared to the PML and STRAIGHT vocoders, in case of male voices. Moreover, the continuous synthesizer was also found to have similar or slightly worse quality than state-of-the-art vocoder in female speaker. Therefore, cNM offers a good alternative method to reconstruct noise than other approaches (for instance, bNM). Further research is necessary to optimize different speech synthesizers with cNM method in order to produce less distortion in the recovered speech signal.

These analysis-synthesis results presented in the current article showed the feasibility of our proposed vocoder, while it is a further step to apply it in statistical parametric speech synthesis, using a bi-directional long-short memory based recurrent neural network. We also plan to use this version of continuous vocoder for voice conversion purposes with a small amount of training data to further improve the perceptual quality of the converted speech. As cNM parameter is not limited only to our vocoder, we try to apply it to other types of modern parametric vocoders (such as Ahocoder [36] as well as PML [12]) to deal with the case of noisy conditions.

Acknowledgments

The research was partly supported by the AI4EU project and by the National Research, Development and Innovation Office of Hungary (FK 124584 and PD 127915). The Titan X GPU used was donated by NVIDIA Corporation. We would like to thank the subjects for participating in the listening test.

References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, USA, pp.373–376, 1996.

[2] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," Speech Communication, vol.51, no.3, pp.1039–1064, 2009.

[3] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," Proc. 9th ISCA Speech Synthesis Workshop (SSW9), Sunnyvale, CA, pp.202–207, 2016.

[4] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech, Language Process., vol.15, no.8, pp.2222–2235, 2007.

[5] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The blizzard challenge," Proc. Blizzard Challenge Workshop, 2008.

[6] A.V. Oord, et al., "WaveNet: A generative model for raw audio," Proc. ISCA Speech Synthesis Workshop, 2016.

[7] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction," Speech Communication, vol.27, no.3, pp.187–207, 1999.

[8] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. Inf. & Syst., vol.E99-D, no.7, pp.1877–1884, 2016.

[9] W. Ping, et al., "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," Proc. 6th International conference on Learning Representations, Vancouver, Canada, pp.1–16, 2018.

[10] M.S. Al-Radhi, T.G. Csapó, and G. Németh, "Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis," Proc. Interspeech, Stockholm, pp.434–438, 2017.

[11] M.S. Al-Radhi, T.G. Csapó, and G. Németh, "Deep recurrent neural networks in speech synthesis using a continuous vocoder," Proc. Speech and Computer (SPECOM), Lecture Notes in Computer Science, Hatfield, England, UK, pp.282–291, 2017.

[12] G. Degottex, P. Lanchantin, and M. Gales, "A log domain pulse model for parametric speech synthesis," IEEE/ACM Trans. Audio, Speech, Language Process., vol.26, no.1, pp.57–70, 2018.

[13] D. Van Compernolle, "Noise adaptation in a hidden markov model speech recognition system," Computer Speech and Language, vol.3, no.2, pp.151–167, 1989.

[14] X. Zhang, K. Demuynck, and H. Van hamme, "Histogram equalization and noise masking for robust speech recognition," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Texas, USA, pp.4578–4581, 2010.

[15] B.A. Mellor and A.P. Varga, "Noise masking in a transform domain," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Minneapolis, pp.87–90, 1993.

[16] G. Kim and P.C. Loizou, "A new binary mask based on noise constraints for improved speech intelligibility," Proc. Interspeech, Chiba, Japan, pp.1632–1635, 2010.

[17] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," EURASIP Journal on Audio, Speech, and Music Processing, vol.38, 2014.

[18] W. Yang and R. Yantorno, "Improvement of MBSD by scaling noise masking threshold and correlation analysis with MOS difference instead of MOS," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Arizona, USA, pp.673–676, 1999.

[19] P.N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," IEEE Signal Processing Letters, vol.20, no.1, pp.102–105, 2013.

[20] T. Drugman and Y. Stylianou, "Maximum voiced frequency estimation: exploiting amplitude and phase spectra," IEEE Signal Process. Lett., vol.21, no.10, pp.1230–1234, 2014.

[21] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis: A unified approach to speech spectral estimation," Proc. International Conference on Spoken Language Processing, Yokohama, Japan, pp.1043–1046, 1994.

[22] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," IEEE Trans. Audio, Speech, Language Process., vol.20, no.3, pp.994–1006, 2012.

[23] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum Approximation (MLSA) filter for speech synthesis," Electronics and Communications in Japan (Part I: Communications), vol.66, no.2, pp.10–18, 1983.

[24] M.S. Al-Radhi, T.G. Csapó, and G. Németh, "Continuous vocoder in feed-forward deep neural network based speech synthesis," Proc. 11th Digital speech and image processing conference, Serbia, pp.1–4, 2017.

[25] N.I. Fisher, Statistical analysis of circular data, Cambridge University, UK, 1995.

[26] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," Speech Communication, vol.67, pp.1–7, 2015.

[27] J. Kominek and A. Black, "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, 2003.

[28] J. Tribolet, P. Noll, B. McDermott, and R.E. Crochiere, "A study of complexity and quality of speech waveform coders," Proc. IEEE International Conference Acoustics, Speech, Signal Processing (ICASSP), Oklahoma, USA, pp.586–590, 1978.

[29] I.M. Kates, "Coherence and speech intelligibility index," The Journal of the Acoustical Society of America, pp.2224–2237, 2005.

[30] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," The Journal of the Acoustical Society of America, vol.125, no.5, pp.3387–3405, 2009.

[31] P.D. Hill, "Kernel estimation of a distribution function," Communications in Statistics - Theory and Methods, vol.14, no.3, pp.605–620, 1985.

[32] E. Parze, "On estimation of a probability density function and mode," The Journal Annals of Mathematical Statistics, vol.33, no.3, pp.1065–1076, 1962.

[33] M.S. Waterman and D.E. Whiteman, "Estimation of probability densities by empirical density functions," International Journal of Mathematical Education in Science and Technology, vol.9, no.2, pp.127–137, 1978.

[34] ITU-R, "Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality," 2001.

[35] S. Huber, "Voice Conversion by modelling and transformation of extended voice characteristics," PhD Thesis, Institute de Recherché et Coordination Acoustique/Musique, France, 2015.

[36] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis," IEEE J. Sel. Topics Signal Process., vol.8, no.2, pp.184–194, 2014.

Mohammed Salah Al-Radhi was born in Basra, Iraq. He got a BSc degree in Computer Engineering at Basra University in 2007, and a MSc degree in Communication Systems at Portsmouth University, UK which was achieved with first honors in 2012 and awarded the MSc top student certificate in 2013. He started from September 2016 to complete his PhD at the Speech Technology and Intelligent Interactions Laboratory in the Budapest University of Technology and Economics. He is working on designing vocoders and acoustic models for statistical speech synthesis. His main interests are the signal processing, speech synthesis, deep learning, acoustic models, and voice conversion.

Tamás Gábor Csapó obtained his MSc in computer science from Budapest University of Technology and Economics (BME), Hungary in 2008. Between 2008-2014, he was a doctoral student at the Speech Technology and Intelligent Interactions Laboratory of BME, where he obtained his PhD degree. In 2007, he was awarded with 1st prize of the National Conference of Scientific Student's Associations, Hungary. He received a CIRE student grant of the Acoustical Society of America in 2010 and was a Fulbright scholar at Indiana University, USA in 2014, where he started to deal with ultrasound tongue imaging. In 2016, he joined the MTA-ELTE Lingual Articulation Research Group, focusing on investigating the Hungarian articulation during speech production. Since 2017, he has two national research projects about ultrasound-based silent speech interfaces. His research interests include speech synthesis, speech analysis, vocoding and ultrasound-based tongue movement analysis.

Géza Németh was born in 1959. He obtained his MSc in electrical engineering, major in Telecommunications at the Faculty of Electrical Engineering of BME in 1983. Also at BME: dr. univ, 1987, PhD 1997. He is an associate professor at BME. He is the author or co-author of more than 150 scientific publications and 4 patents. His research fields include speech technology, service automation, multilingual speech and multimodal information systems, mobile user interfaces and applications. He is the Head of the Speech Technology and Smart Interactions Laboratory of BME TMIT.