9th ISCA Speech Synthesis Workshop
13-15 Sep 2016, Sunnyvale, USA

A Pulse Model in Log-domain for a Uniform Synthesizer

Gilles Degottex[1], *Pierre Lanchantin*[1], *Mark Gales*[1]

[1]Cambridge University Engineering Department, Cambridge, UK
gad27@cam.ac.uk, pkl27@cam.ac.uk, mjfg100@cam.ac.uk

Abstract

The quality of the vocoder plays a crucial role in the performance of parametric speech synthesis systems. In order to improve the vocoder quality, it is necessary to reconstruct as much of the perceived components of the speech signal as possible. In this paper, we first show that the noise component is currently not accurately modelled in the widely used STRAIGHT vocoder, thus, limiting the voice range that can be covered and also limiting the overall quality. In order to motivate a new, alternative, approach to this issue, we present a new synthesizer, which uses a uniform representation for voiced and unvoiced segments. This synthesizer has also the advantage of using a simple signal model compared to other approaches, thus offering a convenient and controlled alternative for future developments. Experiments analysing the synthesis quality of the noise component shows improved speech reconstruction using the suggested synthesizer compared to STRAIGHT. Additionally an experiment about analysis/resynthesis shows that the suggested synthesizer solves some of the issues of another uniform vocoder, Harmonic Model plus Phase Distortion (HMPD). In text-to-speech synthesis, it outperforms HMPD and exhibits a similar, or only slightly worse, quality to STRAIGHT's quality, which is encouraging for a new vocoding approach.
Index Terms: parametric speech synthesis, vocoder, pulse model

1. Introduction

Statistical Parametric Speech Synthesis (SPSS) systems are useful technologies for many applications and can also be a necessary means for communication in case of speech impairment [1]. Even though, current SPSS systems provide a sufficient quality for some applications (e.g. GPS devices in noisy environment), it is still not satisfying for many others (e.g. applications in quiet environments, entertainment industry). Regarding this issue, the vocoder used for reconstructing the waveform from the generated parameters, is critical since it is responsible, together with the features it uses, for a substantial part of the current degradation [2]. The capacity of the vocoder to resynthesize all of the components of the speech signal is obviously important for obtaining all of the perceived characteristics the voice can produce. Otherwise, the vocoder, as well as the SPSS system using it, would be locked on a particular voice quality that might perfectly fit for a specific set of voices, but would systematically fail at reproducing the rest of the voice space. The flexibility of the vocoder's model will play a critical role in this matter. For example, representing the speech signal in a uniform way across time and frequency, e.g. using the same representation for both voiced and unvoiced segments, it allows both smooth and abrupt transitions at different time for different frequency bands. It also avoids discontinuities at both feature and waveform levels, that do not necessarily appear in transients and can impact the quality [5]. It also alleviates the dependency of the SPSS system with respect to a voicing detector, thus, simplifying the learning process [4, 5]. The simplicity of the model is also an important property, which is often neglected. Indeed, complex models also implies complex implementations that are difficult to modify and improve for testing new ideas in a controllable way. Also, over-parametrization of models often lead to intractable tuning issues that depend on very specific expertise and know-how.

STRAIGHT is currently the most used vocoder for SPSS [6, 7], which uses a voicing decision in order to ensure the full randomization of the unvoiced segments, like other vocoders [8, 9]. The noise component in voiced segments is analyzed and reconstructed using an aperiodicity measure. Basically, this measure computes the difference between an upper envelope, which is based on harmonic peaks, and a lower envelope, which is based on spectral valleys [7]. In noisy time-frequency regions of voiced segments, this measure underestimates the noise level because this upper-to-lower difference is always positive and substantial, whereas it should be close to zero in these regions in order to obtain a proper resynthesis of the noise level. Therefore, the noise that should be reproduced in the synthetic waveform tends to be lower than that of the original signal (as shown and illustrated in Sec. 3.1). On the one hand, this underestimation is a *safe* approach for vocoding, since it minimizes the risk of over-randomizing the voiced part of the transients. Indeed, it has been shown that a lack of noise (i.e. leading often to buziness) is preferred over noisiness in the transients [3]. Additionally, this *safe* approach also minimizes the noise generated in creaky voice segments that easily become hoarse if the noise level is overestimated. This overestimation actually occurs in creaky voice since most noise estimators mistake additive noise with randomness of pulse positions. This leads to very high estimated noise level in creaky voice whereas the glottal pulses is actually closer to a Dirac in this mode of phonation [10]. On the other end, by mitigating the noise component, this *safe* approach tends to produce always the same voice quality, a slightly tense and buzzy voice. As mentioned above, this is sort of a deadlock for vocoding, since it eludes the problem of an accurate noise resynthesis that is necessary for a good reconstruction of breathiness and other voice qualities that involve the presence of noise in voiced segments, and ultimately for the overall quality. In other words, for improving the flexibility that vocoders need for covering a bigger range of voice qualities, one way or another, it will be necessary to manage the noise component properly.

Conversely to STRAIGHT, the Harmonic Model + Phase Distortion (HMPD) vocoder uses a uniform representation [5]. The noise that is present in both voiced and unvoiced segments is driven by a Phase Distortion Deviation (PDD) that is used to randomize the phase of the harmonics [5]. Even though HMPD

http://dx.doi.org/10.21437/SSW.2016-35

constitutes an interesting attempt for a uniform model, the synthetic content is limited to harmonic frequencies, which raises the following two issues. Firstly, for mid and high pitch voices, the harmonics are not dense enough with respect to the resolution of the auditory system, so that buziness effects also occur in unvoiced segments, even though the harmonics' phase might be fully randomized. Secondly, no noise can be generated between harmonics, so that voices often lack breathiness, especially falsetto voices, which occurs often in female voices.

In this paper, we want to address the issues above by suggesting a new and simple synthesizer that should reproduce the noisy time-frequency regions of the speech signal more accurately than the two vocoders mentioned above. Since we will be using known features and we suggest only a new synthesis procedure, we use the term *synthesizer* and not *vocoder* in the following. The used signal model, called *Pulse Model in Log-domain* (PML), generates a sequence of wide-band pulses, in spectral domain, similarly to the STRAIGHT vocoder [6, 7] and conversely to HMPD that synthesises harmonics. In both voiced and unvoiced segments, a *pulse* is a morphing between a Dirac function and a short segment of Gaussian noise, followed by the convolution of the Vocal Tract Filter (VTF). Thus, conversely to HMPD, the pulse synthesis can generate spectral content at any frequency, thus, solving HMPD issues, while preserving the uniformity of representation. Obtaining a perceptually meaningful morphing between a Dirac and a specific time segment of noise is far from straightforward. For example, using a traditional additive weighting of the two components in linear domain, the Dirac function will disappear only when the noise masks it. Knowing also that the noise level and Dirac amplitude dependent on two different normalisation, the energy and the sum of the window, respectively, controlling this masking effect is far from obvious. For this reason, as well as the underestimated aperiodicity mentioned above, the Dirac component tends to arise from the noise when using an additive weighting, which often leads to extra buziness effects in current vocoders. From this perspective, even though the traditional source-filter model is well supported by the voice production, it might not be the most practicable way to control the mixture of deterministic and random components of a synthesized speech signal. HMPD alleviates this issue by randomizing the phase of the harmonics proportionally to the PDD feature, which gradually blurs the periodicity. For the suggested PML synthesizer, we aim at preserving this property. We suggest to weight the noise component in the log spectral domain (i.e. multiplication in linear spectral domain, convolution in time domain). The convolution of the Dirac by the noise randomises the Dirac and avoids any possible residual buziness. Additionally, this log-domain formulation leads to a very simple definition of the synthesizer, as shown in the next Section. In this first presentation of PML, we simplified the weighting function to a binary mask. I.e. For each time-frequency bin, the Dirac of each pulse is either left untouched or fully replaced by the corresponding bin of the noise's spectrum. This mask can also be seen as a time-frequency binary voicing decision, which can take any shape and is not limited to time limits (as with voicing decisions) and/or frequency limits (as with a maximum voiced frequency [8]). To limit the differences with the state of the art, this mask is built from the same PDD feature used in HMPD.

We also demonstrate the problem of noise reduction that exists in STRAIGHT and HMPD. The contribution of this paper is thus twofold: i) we show the deadlock that appears with the *safe* approach of STRAIGHT, and ii) we suggest a potential way, through this new synthesizer, that could unlock this situation in the near future. Note that, since we take a more risky, but necessary, approach in this paper, we do not aim at outperforming the state of the art in this first presentation. As it can be understood from above, the development of a full vocoder (features+synthesizer) that will outperform the state-of-the-art vocoders for the majority of voices goes beyond this single paper. We aim at suggesting a synthesizer that offers a simplicity and flexibility that current approaches do not have. In future works, these properties should help to better control the components of the speech signal and help to elaborate new features or techniques that should overcome the current deadlock.

Sec. 2 describes the PML synthesizer in details. Sec. 3.3 first illustrates the current limitation in terms of noise synthesis and then presents results of listening tests for analysis/resynthesis and for parametric text-to-speech synthesis.

2. The PML Synthesizer

The PML synthesis process needs the following features that are illustrated in Fig. 1: i) A fundamental frequency curve $f_0(t)$, which exhibits no voicing decisions. If the provided fundamental frequency contains zeros, these segments can be interpolated linearly between voiced segments, and extrapolated at the beginning and end of the signal. ii) The VTF response $V(t, \omega)$, which is assumed to be minimum phase. iii) A mask $M(t, \omega)$ in the time-frequency space, which is equal to 0 for deterministic regions and 1 for noisy regions. In this work, we derived this mask from the Phase Distortion Deviation (PDD) $\text{PDD}(t, \omega)$, which has been previously used for phase randomization in HMPD [5] and for other applications [11, 12].

2.1. Mask computation

For the first presentation of this model, we chose a very simple approach for computing this mask. Future works might focus on more elaborated strategies. The mask is simply a thresholded version of the PDD measurement. In [5], it is shown that the measurement of phase variance saturates when the variance increases. Consequently, a threshold of 0.75 was used to force the variance to higher values in order to ensure the proper randomization of the noise segments. In this work, we used the same threshold for building the mask:

$$M(t,\omega) = \begin{cases} 0 & \text{PDD}(t,\omega) \leq 0.75 \\ 1 & \text{PDD}(t,\omega) > 0.75 \end{cases} \quad (1)$$

Note that the PDD computation is based on differences between harmonics' phase. Because the harmonics' phase is normalized by the first one [13, 5], a phase difference occurs only from the 2nd harmonic and above. Thus, the PDD computation is zero below the 2nd harmonic and as a consequence, the mask is also zero in this frequency band. This implies that the first harmonic is never randomized. This is actually not a problem since, in silences and fricatives, the corresponding amplitude is rather weak so that this sinusoid is actually never perceived. In voiced segments, this sinusoid is almost always present for all voice qualities.

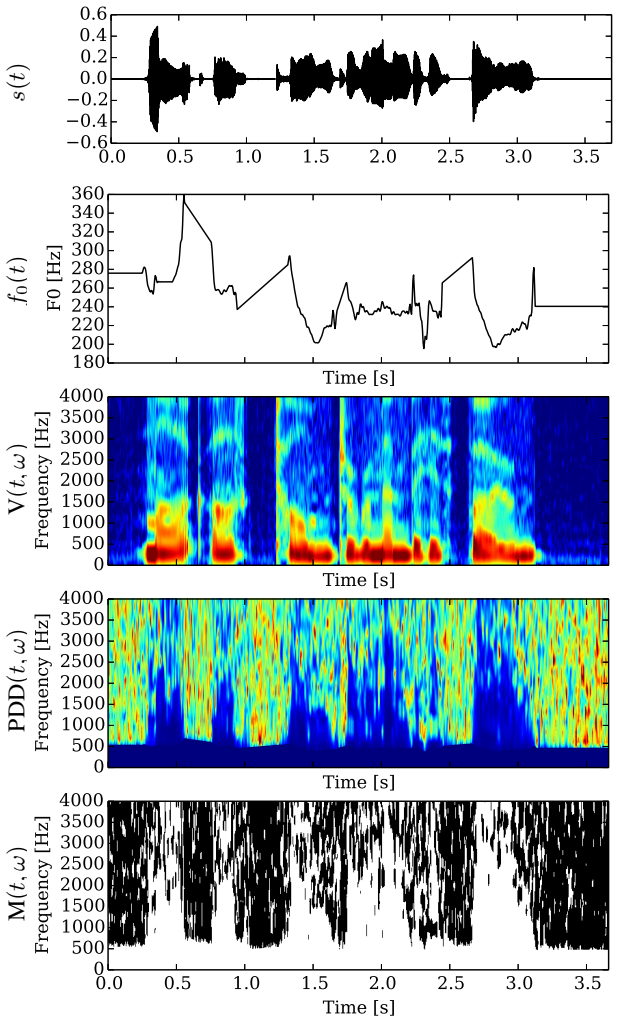

Figure 1: From top to bottom: the waveform used to extract the following elements; The continuous fundamental frequency curve $f_0(t)$; the amplitude spectral envelope $V(t,\omega)$; the Phase Distortion Deviation PDD(t,ω) (a measure of phase randomness. The warmer the colour, the bigger the PDD value and the noisier the corresponding time-frequency region); the binary mask $M(t,\omega)$ derived from PDD, which allows to switch the time-frequency content from deterministic (white) to random (black). The features that are necessary for the synthesizer are only: $f_0(t)$, $V(t,\omega)$ and $M(t,\omega)$.

2.2. Signal synthesis

The generation of the waveform follows a pulse-based procedure, similarly to the STRAIGHT vocoder. Short segments of speech signals (roughly the size of a glottal pulse) are generated one after the other and overlapped-add. In both voiced and unvoiced segments, the voice source is made of a morphing between a deterministic impulse and Gaussian noise. This source is then convolved by the Vocal Tract Filter (VTF) response.

We first generate a sequence of pulse positions t_i according to $f_0(t)$, all along the speech signal:

$$t_{i+1} = t_i + 1/f_0(t_i) \tag{2}$$

with $t_0 = 0$. Then, we suggest to model the speech signal around each instant t_i according to the following simple formula:

$$S_i(\omega) = e^{-j2\pi t_i} \cdot V(t_i,\omega) \cdot N_i(\omega)^{M(t_i,\omega)} \tag{3}$$

where $N_i(\omega)$ is the Fourier transform of a segment of Gaussian noise starting at $\frac{t_{i-1}+t_i}{2}$ and finishing at $\frac{t_i+t_{i+1}}{2}$, which central instant t_i is re-centered around 0 (to avoid doubling the delay $e^{-j2\pi t_i}$ for the noise in $S_i(\omega)$). In order to obtain a proper noise normalisation, $N_i(\omega)$ is normalized by its energy.

To better understand the elements involved in this model, we can have a look at its log-domain representation:

$$\begin{aligned} lS_i(\omega) = \; & \overbrace{-j2\pi t_i}^{\text{Position}} + \overbrace{\log|V(t_i,\omega)|}^{\text{Amplitude}} + \overbrace{j\angle V(t_i,\omega)}^{\text{Minimum phase}} \\ & + \underbrace{M(t_i,\omega)}_{\text{Noise extent}} \cdot \Big(\underbrace{\log|N_i(\omega)|}_{\text{Noise amplitude}} + \underbrace{j\angle N_i(\omega)}_{\text{Phase randomi.}} \Big) \end{aligned} \tag{4}$$

The *Position* defines the overall position of the voice source. This corresponds to the position of the Dirac delta of the deterministic source component. The *Amplitude* defines the amplitude spectral envelope of the resulting segment of speech. The *Minimum phase* is built from the *Amplitude* through the Hilbert transform in order to delay the energy of the pulse, as resonators do. The *Noise extent* provides the means to switch between deterministic or random voice source at any time-frequency point. For $M(t,\omega) = 1$, the *Noise amplitude* will mainly correct the *Amplitude* in order to account for the difference between deterministic and noise normalisation (sum and energy, respectively). This ensures that the noise amplitude is always aligned on the given *Amplitude* spectral envelope $|V(t,\omega)|$. Note that this would still holds for a continuous $M(t_i,\omega)$ (instead of binary one). With $M(t,\omega) = 1$, the *Phase randomization* will also blur the phase of the Dirac delta and replace it by that of noise. In terms of model control, PML drastically simplifies the handling of the noise in the speech signal. Firstly, its amplitude is controlled by $|V(t,\omega)|$, like the deterministic content. Thus, the extent of noise does not change the perceived amplitude, it basically changes only the nature of the phase. Secondly, masking effects and their difficult mastery, as seen in the traditional source-filter model and discussed above, are avoided. Thirdly, the extent of noise is always a value in $[0,1]$. This suggested model is still basically a source-filter model, but the addition is in the log-domain instead of the linear domain, thus, explaining the chosen name PML.

The pulses around each t_i are finally summed for reconstructing the complete signal:

$$s(t) = \sum_{i=0}^{I-1} \mathcal{F}^{-1}\Big(S_i(\omega)\Big) \tag{5}$$

where I is the number of pulses in the synthesized signal.

This description needs a few complementary technical remarks. Firstly, in the implementation, $S(\omega)$ is obviously replaced by its discrete counterpart. A DFT size of 4096 was used for the following experiments. For reason of efficiency, instead of using a DFT size that covers the whole synthetic signal, the DFT used for each pulse can be reduced in order to cover only an interval around each instant t_i (e.g. 2 periods before t_i and 50ms after t_i in order to leave space for the VTF impulse response to decay without being cut). Secondly, the signal has no energy before $\frac{t_{i-1}+t_i}{2}$ since $V(t_i,\omega)$ is assumed to be minimum phase. Because of the delays introduced by $V(t_i,\omega)$,

there are, however, energy after $\frac{t_i+t_{i+1}}{2}$. This does not create, however, any energy issue since the energy is only delayed and each pulse synthesises an independent spectral content from the other pulses. In other words, because there is no redundancy in the synthesis process, conversely to the inverse STFT process, there is no need to compensate for any windowing effect. One can also note that there is no ad hoc tunning parameter, except for the threshold of 0.75, which actually depends on the used noise feature, here PDD, but not on the signal model itself. In terms of computational efficiency, the process basically needs only 2 FFT per pulse. One FFT for computing $N_i(\omega)$, which needs a specific duration for each t_i ($(t_{i+1} - t_{i-1})/2$), and one FFT^{-1} for computing the time domain signal. If not pre-computed, the computation of the minimum phase of the VTF $\angle V(t_i, \omega)$ from a given amplitude envelope requires also 2 extra FFT per pulse. This is clearly efficient enough for allowing real-time synthesis.

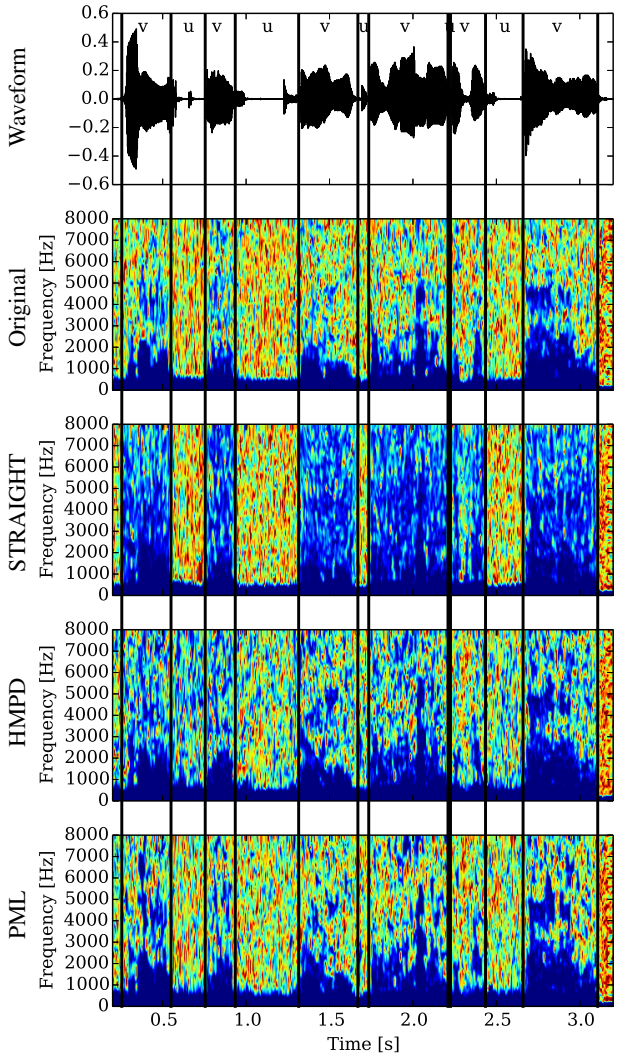

Figure 2: An example of PDD measurements computed from: an original recording and the analysis/resynthesis of STRAIGHT, HMPD and PML (top to bottom). The vertical lines show the voiced/unvoiced transitions used by STRAIGHT. Voiced and Unvoiced segments are annotated by 'v' and 'u', respectively.

2.3. Some important properties for speech signals

It is also worth mentioning the following properties that the suggested model satisfies:

If $M(t, \omega) = 0 \ \ \forall\omega, \forall t$, (3) reduces to:

$$S_i(\omega) = e^{-j2\pi t_i} \cdot V(t_i, \omega) \tag{6}$$

whose corresponding time signal is basically the impulse response of the filter delayed at the pulse position t_i. In this case the signal is thus fully deterministic.

If $M(t, \omega) = 1 \ \ \forall\omega, \forall t$, (3) reduces to:

$$S_i(\omega) = e^{-j2\pi t_i} \cdot N_i(\omega) \cdot V(t_i, \omega) \tag{7}$$

whose corresponding time signal is a filtered noise segment. After summing the terms $S_i(\omega)$, this corresponds to a concatenation process of coloured Gaussian noise segments into a continuous noise signal (the last noise sample of the pulse i is the sample before the first sample of the pulse $i+1$). Thus, no periodicity appears in this noise, even though the synthesis is driven by a continuous $f_0(t)$. In this case, $f_0(t)$ influences only the time resolution of the dynamic noise filtering through the size of the noise segments $(t_{i+1} - t_{i-1})/2$. For f_0 values of 70Hz, a worst case scenario, this still allows to change the noise's colour each 14ms.

3. Experiments

3.1. Noise reconstruction

In this first sub-section, we numerically show the current problem that occurs with the reconstruction of the noise component in two state-of-the-art vocoders (STRAIGHT and HMPD), as discussed in the introduction, and the case of the suggested vocoder based on the PML synthesizer.

Using each 3 vocoder, we first analysed and resynthesized audio samples (i.e. without any statistical modelling) for 6 different English voices [14, 15, 16] (3 females and 3 males; 2 females and 2 males voices at 32kHz sampling rate and 1 female and 1 male voice at 16kHz; 4 American and 2 British). Then, we computed the PDD on the resulting resynthesized signals in order to measure how well the signal randomness is reproduced by each vocoder. Fig. 2 shows an example of this PDD computation over analysis/resynthesis. In unvoiced segments, one can see that the randomness is pretty well reconstructed by all vocoders, except for HMPD. This is expected, since HMPD can reproduce noise only at harmonic frequencies. In voiced segments, the PDD measure over STRAIGHT analysis/resynthesis seems lower than that from the original signal. On the contrary, the PDD measure over PML analysis/resynthesis shows a more accurate reconstruction of the noise extent.

This observation is supported by the estimated distributions of PDD values in the voiced segments shown in Fig. 3. These distributions are computed using 100 samples for each of the 6 voices. The four distributions exhibit basically 2 modes, a small one close to zero and a larger one between 0.5 and 1.5, which roughly correspond to deterministic and noisy time-frequency regions, respectively. Firstly, one can note that the lower mode of the PML's distribution is clearly higher than the others. This is due to the mask that forces the PDD values below 0.75 to zero. Secondly, and more importantly, the higher mode of the distribution corresponding to STRAIGHT's PDD is clearly lower than that of the original signal ($\sim$0.5 instead of $\sim$1.2). Moreover, this mode is below 0.75 for STRAIGHT, whereas it is above this threshold for the original signal, even though

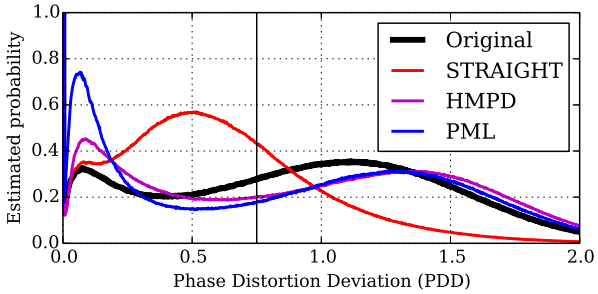

Figure 3: Estimated distributions of PDD measures over analysis/resynthesis using 3 vocoders and the PDD measure on the original speech signals. The vertical line illustrates the threshold of 0.75 used for building the mask in the PML synthesizer.

it was shown that values below this threshold could not lead to the reconstruction of the perceived characteristics of a noise [5]. This demonstrates the reduction of the noise component of STRAIGHT synthesis, as discussed in the introduction. On the contrary, PML better reproduces the higher mode of the original distribution, which should lead to a better reconstruction of noisy components in voiced segments.

3.2. Analysis/Resynthesis quality

In this experiment, we wanted to assess the quality of the analysis/resynthesis of the 3 vocoders, before any use in statistical modelling. For each sound, the corresponding resyntheses from the 3 vocoders used the same amplitude spectral envelope (that of STRAIGHT) and the same $f_0(t)$ curve (that of REAPER [17]). Only the noise features differed, i.e. aperiodicity for STRAIGHT and PDD for HMPD and PML. STRAIGHT used the voicing decision given by REAPER. To carry out this test, we used a Mean Opinion Score (MOS) listening test through a web interface. Each person taking the test had to grade 4 sounds against a reference, where the four sounds where composed of either an analysis/resynthesis using the 3 vocoders or the reference sound itself [18]. Each listener repeated this task for 6 random sentences taken among 100 resyntheses for each of the 6 voices used in the previous experiment. The listening test was advertised on Amazon Mechanical Turk [19, 20] where workers took the test for a small reward. 51 listeners took the test properly and the aggregated results are shown in Fig. 4.

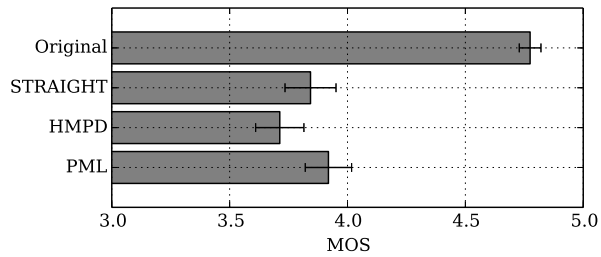

Figure 4: Mean Opinion Scores (MOS) about the analysis/resynthesis quality of 3 vocoders over 6 voices (with the 95% confidence intervals).

From these results, one can see that the quality provided by PML is better than that of HMPD and the confidence interval of STRAIGHT clearly overlaps with those of HMPD and PML. In previous results [5], HMPD's quality was reported to be better than STRAIGHT, which contradicts the results of this test. After inspection of the resynthesized signals, it seems that HMPD struggles in reproducing the creaky voice segments present in the 6 voices of this test. English and mainly American voices, which exhibit a high degree of creaky segments, have been used in this present test. Thus, the degradation in these segments might have been underestimated in the previous tests of HMPD that used a different a set of voices with less creakiness. Because PML synthesises wide-band pulses and not harmonics, it seems to better manage creaky segments than HMPD. We can also conclude that the suggested PML synthesizer provides a similar quality compared to STRAIGHT, while solving the limitations of HMPD mentioned in the introduction and keeping the uniform representation.

A subset of the resyntheses can be found at: `http://gillesdegottex.eu/LT/DemoPMPDResynth/`

3.3. Text-to-speech (TTS) parametric synthesis

For this experiment, we trained HTS-DNN systems for the 3 different vocoders on the 6 voices used above. For each voice, an HTS system [21] was first trained using five-state, left-to-right, no-skip hidden semi-Markov models (HSMMs [21]). Each observation vector consisted of 60 Mel-cepstral coefficients [22], log f_0 values, and 60 Mel-cepstral aperiodicity coefficients or 60 Mel-cepstral PDD coefficients, depending on the vocoder's need, together with the first and second derivatives, extracted every 5ms. Since the aperiodicity is a real-valued spectral measure, like the amplitude spectrum, the basic idea of the Mel-cepstral aperiodicity is to compress the aperiodicity exactly like the amplitude spectrum. This compression technique has two advantages. Firstly, the dimensionality does not depend on the sampling rate of the waveform, conversely to the band aperiodicity. Secondly, high orders can be used (here 59, whereas it is fixed to 24 bands aperiodicity for a 32kHz sampling rate), thus, allowing a statistical model with higher resolution. For this work, this strategy minimizes the impact of the feature compression issue on the studied subject. More importantly, it allows a fair comparison between the TTS systems using STRAIGHT and those using HMPD and PML by using the same dimensionality for the noise feature. For the 6 systems trained for STRAIGHT, a multi-space probability distribution (MSD) [23] was used to model log f_0 sequences consisting of voiced and unvoiced observations (taken from REAPER[17]). For the 6 systems trained for HMPD and PML, no MSD was used since the $f_0(t)$ is continuous. The rest of the topology of the HMM models and systems was similar to the one used for the Nitech-HTS system ([24]). The resulting systems provided state-aligned labels used for training Deep Neural Networks (DNN) in order to improve the features prediction. The used DNN pipeline is exactly the same as the DNN baseline used in [25]. 592 binary and 9 numerical features were derived from the questions used in the HTS systems. The output features were exactly the same as the ones used for the HTS systems. Input features were normalised to [0.01, 0.99] and output features were normalised to zero mean and unit variance. The DNN topology was made of 6 hidden layers of 1024 units. Further details about the learning process can be found in [25].

In order to compare the vocoders and assess their impact on TTS, we carried out a Comparative Mean Opinion Score (CMOS) listening test. Using the systems described above, we synthesized 142 sentences for each of the 6 voices using the duration models of the HTS systems and the features predicted from the DNN systems. Common duration were used between the vocoders, as well as $f_0(t)$ curves and amplitude spectra in order to remove the impact of the prosody and the

influence of the amplitude modelling, which is not the subject of this work. The systems trained for STRAIGHT were used to build these common features ($f_0(t)$ was then linearly interpolated for HMPD and PML). Each listener taking the test assessed the 3 pairs of vocoder combinations for 8 random sentences among the 142x6=852 synthesized sentences [26]. Again, workers from Amazon Mechanical Turk were asked to take the test for a small reward. 53 listeners took the test properly and the aggregated results are shown in Fig. 5. From this figure, one can see that both STRAIGHT and PML outperform HMPD. According to this result and that of

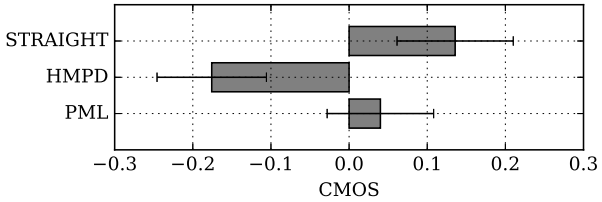

Figure 5: Comparative mean opinion scores (CMOS) for 3 vocoders using HTS-DNN systems over 6 voices (with the 95% confidence intervals).

the previous test, it seems clear that PML solves the major drawbacks of HMPD, while using the same features in the statistical model, while preserving the uniformity of representation between voiced and unvoiced segments and using an even simpler synthesis technique. The confidence intervals between STRAIGHT and PML clearly overlap. However, a strong trend favours the STRAIGHT vocoder. Nevertheless, with regard to the *safe* approach taken by STRAIGHT, as discussed in the introduction, which eludes the difficulty to properly resynthesize the noise component in voiced segments, this result is quite encouraging for future development of better masks or noise control based on PML.

A subset of the syntheses can be found at `http://gillesdegottex.eu/LT/DemoPMPDTTS/`

4. Conclusions

The contribution of this paper was twofold. Firstly, we have shown the noise reconstruction problem that is present in state-of-the-art vocoders and we discussed the limitations that it implies in synthesis of voice qualities and the overall improvement of the vocoders' quality for SPSS technologies. Secondly, we suggested a very simple signal model for a new synthesizer called PML, in order to suggest a new approach to noise synthesis for addressing the issue above.

This synthesizer was shown to better reconstruct the noisiness of the speech signal, compared to STRAIGHT and HMPD vocoders, thus, offering an encouraging alternative for future works in this new approach. In terms analysis/resynthesis quality, this PML synthesizer outperformed the HMPD vocoder, while preserving a uniform time-frequency representation for both voiced and unvoiced segments. Even though PML was found to have only similar or slightly worse quality than STRAIGHT in a text-to-speech experiment, the uniformity, the flexibility and the simplicity of the suggested PML synthesizer is quite encouraging for future developments, in order to tackle the current limitations of voice quality reconstruction.

Future works will focus on continuous masks for morphing the deterministic content into noise. Because it relies on a harmonic model, the used PDD feature, which is currently used for building this mask, has also some limitations that should be addressed, especially in creaky voice segments.

5. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 655764. The research for this paper was also partly supported by EPSRC grant EP/I031022/1 (Natural Speech Technology).

6. References

[1] C. Veaux, J. Yamagishi, and S. King, "Using hmm-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders." in *Proc. Interspeech*, 2012.

[2] G. Henter, T. Merritt, M. Shannon, C. Mayo, and S. King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *Proc. Interspeech*, vol. 15, 2014, pp. 1504–1508.

[3] J. Latorre, M. J. F. Gales, S. Buchholz, K. Knill, M. Tamurd, Y. Ohtani, and M. Akamine, "Continuous f0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?" in *Proc. ICASSP*, 2011, pp. 4724–4727.

[4] K. Yu and S. Young, "Continuous f0 modeling for HMM-based statistical parametric speech synthesis," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 5, pp. 1071–1079, 2011.

[5] 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. 2014, no. 38, 2014.

[6] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[7] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *MAVEBA*, 2001.

[8] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, 2014.

[9] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4230–4234.

[10] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech & Language*, vol. 28, no. 5, pp. 1233 – 1253, 2014.

[11] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, "The importance of phase on voice quality assessment," in *Proc. Interspeech*. Singapore: International Speech Communication Association (ISCA), September 2014.

[12] G. Degottex and N. Obin, "Phase distortion statistics as a representation of the glottal source: Application to the classification of voice qualities," in *Proc. Interspeech*. Singapore: International Speech Communication Association (ISCA), September 2014.

[13] I. Saratxaga, I. Hernaez, M. Pucher, and I. Sainz, "Perceptual Importance of the Phase Related Information in Speech," in *Proc. Interspeech*. ISCA, 2012.

[14] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. ISCA Speech Synthesis Workshop*, 2003, pp. 223–224, http://www.festvox.org/cmu_arctic.

[15] M. Cooke, C. Mayo, and C. Valentini-botinhao, "Intelligibilityenhancing speech modifications: the Hurricane Challenge," in *in Proc. Interspeech*, 2013.

[16] The Speech Synthesis Special Interest Group, "The Blizzard Challenge 2016 [Online]," http://www.synsig.org/index.php/Blizzard_Challenge_2016/, 2016.

[17] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR [Online]," Github: https://github.com/google/REAPER, 2015.

[18] The ITU Radiocommunication Assembly, "ITU-R BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems," ITU, Tech. Rep., 2003.

[19] M. K. Wolters, K. B. Isaac, and S. Renals, "Evaluating speech synthesis intelligibility using Amazon Mechanical Turk," in *Proc. 7th Speech Synthesis Workshop (SSW7)*, 2010, pp. 136–141.

[20] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Proc. Interpseech*, 2011, pp. 3053–3056.

[21] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[22] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137–140.

[23] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synthesis Workshop*, 2002.

[24] H. Zen, T. Toda, M. Nakamura, and T. Tokuda, "Details of the nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[25] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, 2015, pp. 4460–4464.

[26] The ITU Radiocommunication Assembly, "ITU-R BS.1284-1: En-general methods for the subjective assessment of sound quality," ITU, Tech. Rep., 2003.