INTERSPEECH 2009 BRIGHTON

A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis

Thomas Drugman [1], *Geoffrey Wilfart* [2], *Thierry Dutoit* [1]

[1] TCTS Lab, Faculté Polytechnique de Mons, Belgium
[2] Research and Development, Acapela Group, Mons, Belgium

thomas.drugman@fpms.ac.be

Abstract

Speech generated by parametric synthesizers generally suffers from a typical buzziness, similar to what was encountered in old LPC-like vocoders. In order to alleviate this problem, a more suited modeling of the excitation should be adopted. For this, we hereby propose an adaptation of the Deterministic plus Stochastic Model (DSM) for the residual. In this model, the excitation is divided into two distinct spectral bands delimited by the maximum voiced frequency. The deterministic part concerns the low-frequency contents and consists of a decomposition of pitch-synchronous residual frames on an orthonormal basis obtained by Principal Component Analysis. The stochastic component is a high-pass filtered noise whose time structure is modulated by an energy-envelope, similarly to what is done in the Harmonic plus Noise Model (HNM). The proposed residual model is integrated within a HMM-based speech synthesizer and is compared to the traditional excitation through a subjective test. Results show a significative improvement for both male and female voices. In addition the proposed model requires few computational load and memory, which is essential for its integration in commercial applications.

Index Terms: HMM-based speech synthesis, residual modeling, Deterministic plus Stochastic model

1. Introduction

Statistical parametric speech synthesizers have recently shown their ability to produce natural-sounding voices [1],[2]. Compared to other state-of-the-art techniques, they present the advantage of being flexible while requiring a low footprint. This makes them particularly suited for small devices. On the other side, their main drawback lies in the buzziness of the delivered quality, as typically encountered in LPC-like vocoders. This can be mostly explained by the parametrical representation the synthesizer relies on. While methods capturing the spectral envelope have nowadays reached maturity, there still is a lot to be gained in finding a suited modeling of the excitation.

The traditional excitation used by HMM-based speech synthesizers is either a pulse train or a white noise, during voiced and unvoiced segments respectively. In order to enhance this model, some approaches have been proposed in the literature. In [3], Yoshimura et al. integrated the Mixed Excitation (ME) coding method. In this framework, the excitation is obtained using a multi-band mixing model, containing both periodic and aperiodic contributions, and controled by bandpass voicing strengths. In a similar way, Maia et al. [4] made use of high-order filters to obtain these components. In [5] and [6], Cabral et al. suggested the integration of a Liljencrants-Fant waveform ([7]) as a modeling of the glottal source. In [8] we proposed the use of a codebook of pitch-synchronous residual frames to construct the voiced excitation. All these techniques tend to relatively reduce the produced buzziness, and therefore improve the overall quality.

The main motivation of this paper arises from the unsatisfying results we obtained on female voices in [8]. Indeed while this latter method led to a convincing improvement over the traditional excitation for the male speakers, subjective tests were more mitigated for the female voices. For this, we propose an adaptation of the Deterministic plus Stochastic Model (DSM) for the residual signal. The most famous example of DSM is the Harmonic plus Noise Model (HNM) which already showed its high-quality parametrization of speech [9],[10]. According to this model, speech is composed of a low-frequency harmonic structure and a high-frequency noise, assumed to principally model the turbulences of the glottal airflow. The spectrum is then divided into two bands delimited by the so-called *maximum voiced frequency* F_m. Our idea in this paper is to adopt a similar approach on the residual signal. While the stochastic part may stay unchanged, some modifications have to be brought to the deterministic component. For this, similarly to [8], we propose to model the low-frequency contents (below F_m) by decomposing pitch-synchronous residual frames on an orthonormal basis obtained by Principal Component Analysis (PCA). The resulting DSM then consists of a compact representation of the residual, requiring few computational load and memory, which make it suited for its integration within small devices.

The paper is structured as follows. Section 2 details the way our pitch-synchronous residual frames are obtained and how they are modeled by the Deterministic plus Stochastic Model (DSM). Section 3 presents the incoporation of the DSM in our HMM-based speech synthesizer. The traditional and proposed excitations are then compared by a subjective test in Section 4. Finally Section 5 concludes.

2. A Deterministic plus Stochastic Model of pitch-synchronous residual frames

This Section first describes how the pitch-synchronous residual frames are obtained from speech signals (2.1). These frames have the particularity to be centered on a Glottal Closure Instant (GCI), two period-long and Blackman-windowed. This processing makes them comparable so that they are suited for a common modeling. For this, we apply a DSM (2.2) consisting of (a) a low-frequency deterministic model, based on a PCA decomposition (2.2.2), and (b) a high-frequency modulated noise component (2.2.3).

2.1. Pitch-synchronous residual frames

The workflow for obtaining the pitch-synchronous residual frames is presented in Figure 1. First a Mel-Generalized Cepstral (MGC) analysis is performed on the speech signals, as these features have shown their efficiency to capture the spectral envelope [11]. As recommended in [2], we used the parameter values $\alpha = 0.42$ ($Fs = 16kHz$) and $\gamma = -1/3$ for the MGC extraction. Residuals are then obtained by inverse filtering. Glottal Closure Instants (GCIs) are then identified by locating the greatest discontinuity in the residual signal as explained in [12]. In parallel, the pitch is estimated using the publicly available Snack Sound Toolkit [13]. The pitch-synchronous residuals are finally isolated by a GCI-centered and two period-long Blackman windowing.

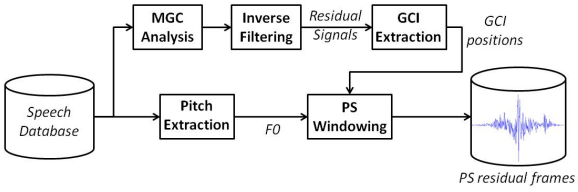

Figure 1: Workflow for obtaining the pitch-synchronous residual frames.

2.2. The proposed Deterministic plus Stochastic Model

As previously mentioned, the DSM consists of the superposition of a deterministic $r_d(t)$ and stochastic $r_s(t)$ components of the residual. These components act in two distinct spectral bands delimited by the frequency F_m, equivalent to the *maximum voiced frequency* in the HNM model. The DSM we applied on the pitch-synchronous residual frames is described here below.

2.2.1. The maximum voiced frequency

The maximum voiced frequency demarcates the boundary between both deterministic and stochastic components. Although some methods have been proposed for its estimation [9], we use in this work a fixed value at 4 kHz, as done in [10] or [14].

2.2.2. The deterministic modeling

In order to model their low-frequency contents, we propose to decompose pitch-synchronous residual frames (as extracted in 2.1) on an orthonormal basis obtained by PCA. For this, a dataset of such frames is constructed, and frames are normalized in both pitch and energy. This step ensures the coherence of the dataset before applying PCA. Assuming the residual signal as an approximation of the glottal source, resampling the residual frames by interpolation and decimation should preserve their shape and consequently their most important features. Similarly to [8], care has to be taken when choosing the number of points for the length-normalization. In order to avoid the appearance of energy holes at synthesis time (occuring if the useful band of the deterministic part does not reach F_m), the pitch value F_0^* for the normalization is such that:

$$F_0^* \leq \frac{F_N}{F_m} \cdot F_{0,min} \tag{1}$$

where F_N and $F_{0,min}$ respectively denote the Nyquist frequency and minimum pitch value for the considered speaker. PCA can now be calculated on the dataset allowing dimensionality reduction and feature decorrelation. The number of retained eigenvectors is chosen such that they explain around 80% of the total dispersion. Through unformal Analysis/Synthesis experiments, this value was reported to give almost inaudible degradations. Besides we observed that once the dataset for PCA computation reached about 10 minutes, extracted eigenvectors remained sensibly unchanged. Figure 2 shows the evolution of the covered relative dispersion with the number of eigenvectors. In this case, using 15 features is sufficient for giving high-quality coding results. For information, the first eigenvector is exhibited in Figure 3. A strong similarity with the LF model [7] can be noticed.

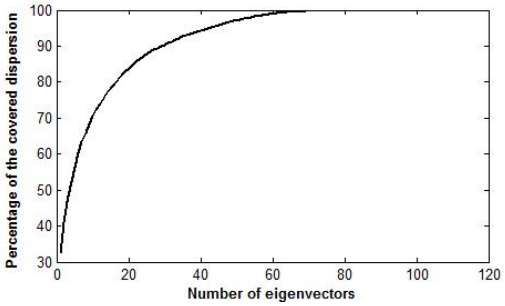

Figure 2: Evolution of the relative dispersion covered with an increasing number of eigenvectors.

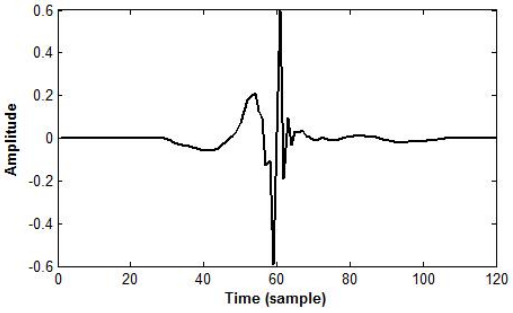

Figure 3: The first eigenvector for speaker SLT.

2.2.3. The stochastic modeling

The stochastic modeling of the residual $r_s(t)$ that we adopted is identical to the noise part in the HNM model [9]. It corresponds to a white Gaussian noise $n(t)$ convolved with an auto-regressive model $h(\tau, t)$ and whose time structure is controled by a parametric envelope $e(t)$:

$$r_s(t) = e(t) \cdot [h(\tau, t) \star n(t)] \tag{2}$$

Although some studies focus on a better modeling of the envelope $e(t)$ [14], we use in this work a pitch-dependent triangular function as suggested in [10]. Furthermore, since F_m is fixed to 4 kHz in our case and that the residual spectrum is almost flat over the whole frequency range, it is reasonable to consider that the auto-regressive model has the same effect on all the frames: it acts as a high-pass filter (beyond F_m) slightly attenuated in the very high frequencies. Consequently $h(\tau, t)$ is computed once and for all for the rest of this work. An example of DSM decomposition is displayed in Figure 4.

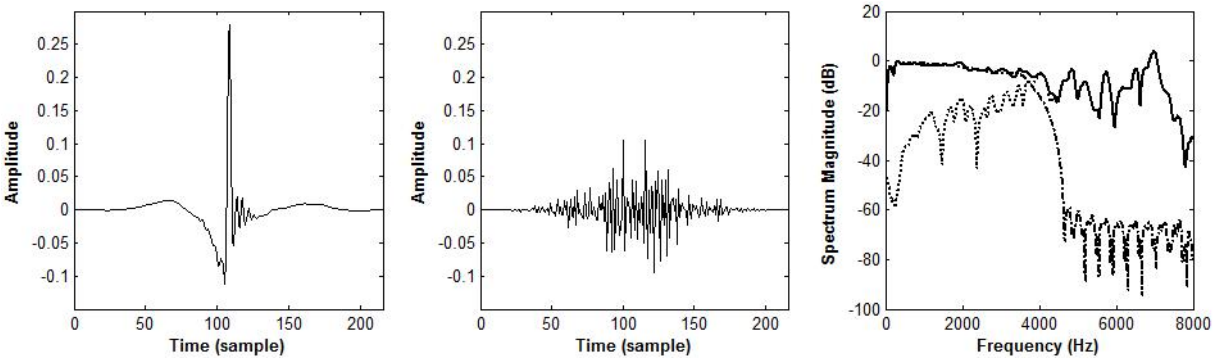

Figure 4: Example of DSM decomposition on a pitch-synchronous residual frame. *Left panel:* the deterministic part. *Middel panel:* the stochastic part. *Right panel:* amplitude spectra of the deterministic part (dashdotted line), the noise part (dotted line) and the reconstructed excitation frame (solid line) composed of the superposition of both components.

2.2.4. The DSM vocoder

A workflow summarizing the DSM vocoder can be found in Figure 5. Inputs are the PCA weights, the pitch and the MGC coefficients. A first version of the deterministic component of the residual $r_d(t)$ is obtained by a linear combination of the eigenvectors. The resulting waveform is resampled such that its length is twice the target pitch period. Following Equation 2, the stochastic part $r_s(t)$ is a white noise modulated by an AR model and multiplied by a triangular envelope centered on the current GCI. Both components are superposed and then overlap-added so as to obtain the residual signal. Note that in the case of unvoiced regions, the excitation consists of a simple white Gaussian noise. The residual is finally the input of the Mel-Log Spectrum Approximation (MLSA) filter to get the speech signal.

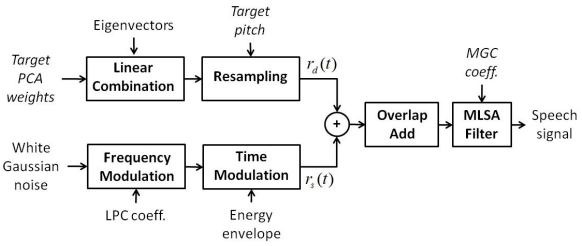

Figure 5: Workflow of the DSM vocoder.

3. Integration in the HMM-based speech synthesizer

The implementation of our HMM-based speech synthesizer relies on the HTS toolkit publicly available in [15]. Some modifications are nonetheless required to incoporate the proposed excitation. For this, three streams of data are considered: one stream for the MGC coefficients, one for the pitch, and one for the PCA weights of the deterministic part. For the last two streams, one needs to distinguish between voiced and unvoiced regions, which led us to adopt a multi-space distribution (MSD), as described in [16]. Following [17], the first derivatives of each stream are included into our model, so that the configuration is the following:

- one single Gaussian distribution with diagonal covariance for MGC coefficients and their derivatives.
- one MSD distribution for pitch,
- one MSD distribution for pitch derivatives,
- one MSD distribution for PCA weights,
- one MSD distribution for PCA weight derivatives.

In each MSD distribution, for voiced parts, parameters are modeled by single Gaussian distributions with diagonal covariance, while the voiced/unvoiced decision is modeled by an MSD weight. At synthesis time, parameters generated from a constrained maximum likelihood algorithm, as described in [17], are fed into our DSM vocoder to produce the synthetic speech. We have observed that, in the generated residuals, the contribution of the first eigenvector clearly dominates the deterministic component. Listening tests confirmed that using eigenvectors of superior ranks led to no audible difference. For this reason, experiments presented in Section 4 were carried out using only the first eigenvector. Under this assumption, excitation is only characterised by the pitch, and the MSD stream of PCA weights may be removed. This leads to a very simple model, in which the voiced excitation essentially (below Fm) consists of a single waveform that is resampled to the target pitch period, requiring almost no computational load, while providing high-quality synthesis.

4. Synthesis experiments

The synthetic voices of five speakers were assessed: AWB (Scottish male), Bruno (French male), Julie (French female), Lucy (US female) and SLT (US female). AWB and SLT come from the publicly available CMU ARCTIC database [18] and about 45 minutes of speech for each were used for the training. Other voices were kindly provided by Acapela Group and were trained on a corpus of around 2 hours. The test consists of a subjective comparison between the proposed and the traditional pulse excitation. For this, 40 people participated to a Comparative Mean Opinion Score (CMOS,[19]) test composed of 20 randomly chosen sentences of about 7 seconds. For each sentence they were asked to listen to both versions (randomly shuffled) and to attribute a score according to their overall preference (see Table 1).

Preference scores can be viewed in Figure 6. A clear improvement over the traditional pulse excitation can be observed for all voices. Compared to the method using a pitch-

Much better	+3
Better	+2
Slightly better	+1
About the same	0
Slightly worse	-1
Worse	-2
Much worse	-3

Table 1: Grades in the CMOS scale. The reference signal used either the traditional or proposed excitation.

synchronous residual codebook proposed in [8], results are almost similar on male speakers, with a minor loss of less than 5% for both AWB and Bruno. On the contrary, the contribution on female voices is much more evident. While only 30% of participants prefered the technique using the codebook for speaker SLT [8], score now reaches more than 90% for the DSM. This trends holds for other female voices. Figure 7 exhibits the average CMOS scores with their 95% confidence intervals. Average values vary between 1 and 1.75, confirming a clear advantage for the proposed technique.

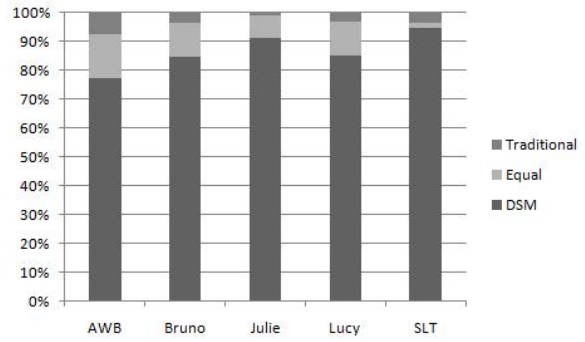

Figure 6: Preference score for the five speakers.

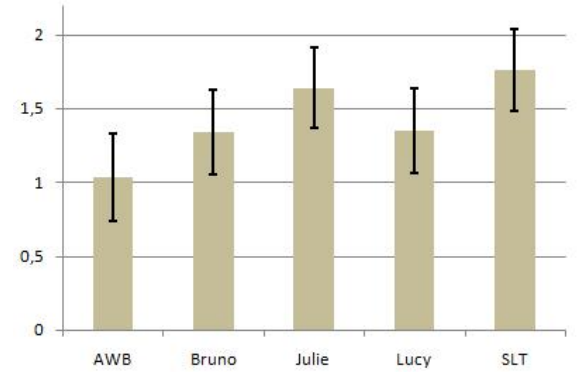

Figure 7: Average CMOS score in advantage of the DSM for the five speakers, together with their 95% confidence interval.

5. Conclusion

This paper proposed an adaptation of the Deterministic plus Stochastic Model for residual signals. This approach is motivated by the need of finding an efficient and compact representation of the excitation, so as to reduce the buzziness produced by parametric speech synthesizers. Through subjective tests, the method was shown to clearly outperform the state-of-the-art excitation for five speakers. In addition the mitigated results we obtained for female voices in a previous work are now undeniably overcome. In the same time, the method preserves the small footprint as well as the weak computational requirements, making it suited for its integration within small devices.

6. Acknowledgments

Thomas Drugman is supported by the "Fonds National de la Recherche Scientifique" (FNRS). Authors also would like to thank the participants of the subjective test.

7. References

[1] Black, A., Zen, H. and Tokuda, K., "Statistical Parametric Speech Synthesis", IEEE ICASSP, pp. 1229-1232, 2007.

[2] Zen, H., Toda, T. and Tokuda, K., "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006", IEICE Trans. on Information and Systems, 2006.

[3] Yoshimura, T., Tokuda, K., Masuko, T. and Kitamura, T., "Mixed-excitation for HMM-based speech synthesis", Eurospeech, pp. 2259-2262, 2001.

[4] Maia, R., Toda, T., Zen, H., Nankaku, Y. and Tokuda, K., "An excitation model for HMM-based speech synthesis based on residual modeling", ISCA SSW6, 2007.

[5] Cabral, J., Renals, S., Richmond, K. and Yamagishi, J., "Towards an Improved Modeling of the Glottal Source in Statistical Parametric Speech Synthesis", ISCA SSW6, 2007.

[6] Cabral, J., Renals, S., Richmond, K. and Yamagishi, J., "Glottal Spectral Separation for Parametric Speech Synthesis", Proc. Interspeech, pp. 1829-1832, 2008.

[7] Fant, G., Liljencrants, J. and Lin, Q., "A four parameter model of glottal flow", STL-QPSR4, pp. 1-13, 1985.

[8] Drugman, T., Wilfart, G., Moinet, A. and Dutoit, T., "Using a Pitch-Synchronous Residual Codebook for Hybrid HMM/frame Selection Speech Synthesis", IEEE ICASSP, 2009.

[9] Stylianou, Y., "Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis", IEEE Trans. Speech and Audio Processing, 9(1):21-29, 2001.

[10] Stylianou, Y., "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification", PhD thesis, Ecole Nationale Superieure des Telecommunications, 1996.

[11] Tokuda, K., Kobayashi, T., Masuko, T. and Imai, S., "Mel generalized cepstral analysis A unified approach to speech spectral estimation", ICSLP, 1994.

[12] Drugman, T. and Dutoit, T., "Glottal Closure and Opening Instant Detection from Speech Signals", submitted to Interspeech 2009.

[13] *[Online]*, "The Snack Sound Toolkit", *http://www.speech.kth.se/snack/*.

[14] Pantazis, Y., and Stylianou, Y., "Improving the modeling of the noise part in the Harmonic plus Noise Model of speech", IEEE ICASSP, 2008.

[15] *[Online]*, "HMM-based Speech Synthesis System (HTS)", *http://hts.sp.nitech.ac.jp/*.

[16] Tokuda, K., Masuko, T., Myiazaki, N. and Kobayashi, T., "Multi-space probability distribution HMM", IEICE Trans. on Information and Systems, vol. E85-D, pp.455-464, 2002.

[17] Tokuda, K., Masuko, T., Yamada, T., Kobayashi T. and Imai, S., "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features", Proc. Eurospeech, 1995.

[18] *[Online]*, "CMU ARCTIC speech synthesis databases", *http://festvox.org/cmu_arctic/*.

[19] Grancharov, V. and Kleijn, W., "Speech Quality Assessment", Springer Handbook of Speech Processing, chap. 5, 2007.