# A Continuous Vocoder using Sinusoidal Model for Statistical Parametric Speech Synthesis

Mohammed Salah Al-Radhi[1], Tamás Gábor Csapó[1,2], Géza Németh[1]

[1]Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Budapest, Hungary
[2]MTA-ELTE Lendület Lingual Articulation Research Group, Budapest, Hungary
{malradhi,csapot,nemeth}@tmit.bme.hu

**Abstract.** In our earlier work in statistical parametric speech synthesis, we proposed a source-filter based vocoder using continuous F0 (contF0) in combination with Maximum Voiced Frequency (MVF), which was successfully used with deep learning. The advantage of a continuous vocoder in this scenario is that vocoder parameters are simpler to model than conventional vocoders with discontinuous F0. However, our vocoder lacks some degree of naturalness and still not achieving a high-quality speech synthesis compared to the well-known vocoders (e.g. STRAIGHT or WORLD). Previous studies have shown that human voice can be modelled effectively as a sum of sinusoids. In this paper, we firstly address the design of a continuous vocoder using sinusoidal synthesis model that is applicable in statistical frameworks. The same three parameters of the analysis part from our previous model have been also extracted and used for this study. For refining the output of the contF0 estimation, post-processing approach is utilized to reduce the unwanted voiced component of unvoiced speech sounds, resulting in a smoother contF0 track. During synthesis, a sinusoidal model with minimum phase is applied to reconstruct speech. Finally, we have compared the voice quality of the proposed system to the STRAIGHT and WORLD vocoders. Experimental results from objective and subjective evaluations have shown that the proposed vocoder gives state-of-the-art vocoders performance in synthesized speech while outperforming the previous work of our continuous F0 based source-filter vocoder.

**Keywords:** Continuous vocoder, Speech synthesis, Sinusoidal model, contF0.

## 1 Introduction

Statistical parametric speech synthesis (SPSS) based text-to-speech (TTS) systems have steadily advanced in terms of naturalness during the last two decades. Even though the quality of synthetic speech is still unsatisfying, the benefits of flexibility, robustness, and control denote that SPSS stays as an attractive proposition. Such a statistical framework is guided by the encoding-decoding (vocoder, which is also called voice coders) concept which is based on parameterization of the speech waveform and reconstruction. Hence, vocoder performance is the most important factor limiting the impact

of overall voice quality in SPSS [1]. Vocoders attempt to produce a decoded signal that sounds like the original speech. Therefore, several approaches based on mathematical and physical models have been suggested to model the overall speech signal.

In recent years, a number of sophisticated source-filter based vocoders have been proposed and extensively used in speech synthesis. Specifically, for example, STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum) vocoder [2] is probably the most used vocoder for SPSS which decomposes signals into spectral envelope, excitation, and aperiodicity parameters. For real-time processing, the computational issue is expensive in STRAIGHT. Furthermore, the Deterministic plus Stochastic Model (DSM) proposed by Drugman et al. [3] is based on a two-band mixed excitation in which the upper band was treated as noise and the lower band was modeled through a set of deterministic waveforms. More recently, a high-quality vocoder named WORLD was developed in [4] to meet the requirements of real-time processing. It consists of three modern algorithms for estimating speech parameters (F0 contour, spectral envelope, and excitation signal used as a band aperiodicity), and one synthesis algorithm.

In our earlier work, we proposed a computationally feasible residual-based source filter vocoder [5], using a continuous F0 model [6], and MVF [7]. In this method, the voiced excitation consisting of pitch synchronous PCA residual frames is low-pass filtered while the unvoiced part is high-pass filtered according to the MVF contour as a cutoff frequency. In [8], we further control the time structure of the high-frequency noise component by estimating a suitable time True envelope. Continuous vocoder was especially successful for modelling speech sounds with mixed excitation.

Sinusoidal vocoder is an alternative category for the source-filter model of speech and has been successfully applied to a broad range of speech processing problems such as speech modification and conversion. Sinusoidal modeling can be characterized by the amplitudes, frequencies, and phases of the component sine waves; and synthesized as the sum of a number of sinusoids that can generate high quality speech. Concisely, voiced speech can be modeled as a sum of harmonics (quasi periodic) spaced at F0 with instantaneous phases, whereas unvoiced speech can be represented as a sum of sinusoids with random phases [9].

Various sinusoidal model formulations have been discussed in the literature. In particular, Harmonic plus Noise Model (HNM) was developed in [10] and has shown the capability of providing high-quality copy synthesis and prosodic modifications. Based on time-varying frequency, HNM decomposes speech into deterministic lower band where the signal is modeled as a sum of harmonically related sinusoids and stochastic upper band where the signal is modeled by colored noise. Another sinusoidal based speech vocoder is being developed by Degottex and Stylianou [11] in which an adaptive Quasi-Harmonic vocoder (aQHM) and Adaptive Iterative Refinement (AIR) method combined as an intermediate model to iteratively minimize the mismatch of harmonic frequencies. Hence, the full system is called aHM-AIR. Similarly, Perception based Dynamic sinusoidal Model (PDM) and Harmonic Dynamic Model (HDM) have been proposed in [12] and have both been applied during analysis and synthesis to be modelled in hidden Markov models (HMM) based speech synthesis.

Thus, from a point of view of either objective or subjective measures, sinusoidal vocoders were preferred in terms of quality. However, these models have usually more parameters (each frame has to be represented by a set of frequencies, amplitude, and phase) than in the source-filter models. Consequently, more memory would be required to code and store the speech segments. Although some experiments have been made to use either an intermediate model [11] or intermediate parameters (regularized cepstral coefficients) [12] to overcome these issues, the computational complexity of SPSS can be quite high once additional algorithms are including [1].

By keeping the number of our vocoder parameters unchanged [8], which are simpler to model than traditional vocoders with discontinuous F0, the goal of this paper is two-fold. The first one is to consider whether a different synthesis technique based on sinusoidal approach can more accurately synthesize speech in a continuous vocoder. Second, an experimental comparison has been made between synthetic speech generated using source-filter continuous vocoder and the proposed one. Besides, the estimated contF0 contours is smoothed by a post-processing phase to eliminate octave errors and isolated glitches. The rest of this paper is structured as follows: Sect. 2 introduced the baseline vocoder. The new form of the continuous vocoder is presented in Sect. 3. Experiment design and our evaluations are discussed in Sect. 4. Finally, Sect. 5 concludes the contributions of this paper.

## 2 Continuous vocoder: Baseline

The baseline vocoder is based on our previous work [8]. During the analysis phase, F0 is calculated on the input waveforms by the open-source implementation[1] of a simple continuous pitch tracker [13]. In regions of creaky voice and in case of unvoiced sounds or silences, this pitch tracker interpolates F0 based on a linear dynamic system and Kalman smoothing. After this step, MVF is calculated from the speech signal using the MVF_Toolkit[2], resulting in the MVF parameter [14]. In the next step 24-order Mel-Generalized Cepstral analysis (MGC) [15] is performed on the speech signal with alpha=0.42 and gamma=-1/3. In all steps, 5 ms frame shift is used. The results are the contF0, MVF and the MGC parameter streams. Finally, we perform Principal Component Analysis (PCA) on the pitch synchronous residuals as shown in our earlier study [5].

During the synthesis phase of the baseline system, voiced excitation is composed of PCA residuals Overlap-Added (OLA) pitch synchronously, depending on the continuous F0. After that, 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 used. True time envelope of the PCA residual has been applied in order to further control the time structure of the high-frequency component in the excitation and noise parts [8]. Voiced and unvoiced excitation is then added together. Finally, MGLSA filter is used to synthesize speech from the excitation and the MGC parameter stream [16].

[1] https://github.com/idiap/ssp

[2] http://tcts.fpms.ac.be/~drugman/files/MVF.zip

## 3 Proposed methodology

This section will show our main contributions of improving the latest version of the continuous vocoder [8]. We illustrate the basic principles of the proposed method by firstly smoothing the contF0 contour to reduce contF0 estimation error, and secondly introducing the novelty of a continuous synthesizer.

### 3.1 Smoothing of the contF0 contour

It is common practice for pitch values to have a small amount of noise variation if they are not estimated well; thereby often leads to extra buzziness or undependable feature measurements. Therefore, a smoothing stage is necessary in improving the quality of the pitch estimation. A variety of smoothing techniques have already been used to smooth the tracked F0 contour and refine the voiced/unvoiced regions. A dynamic programming approach that used in RAPT [17], linear smoothing that use some kind of low-pass filter, and median (nonlinear non-recursive digital filter) approach which utilized in [18] are the most successful smoothing techniques for pitch tracking.

In this paper, we propose to combine two smoothing steps in contF0: 1) median filter using 0.1s window to ignore isolated outliers while preserving both the fine-grained variations and the sharpness of true step transitions; 2) linear smoother (zero-phase filtering) with Hanning window is applied to remove higher-frequency resonance effects and hence suppress the noisiness of the measurement. An example of contF0 estimation made with smoothing technique is shown in Fig. 1.

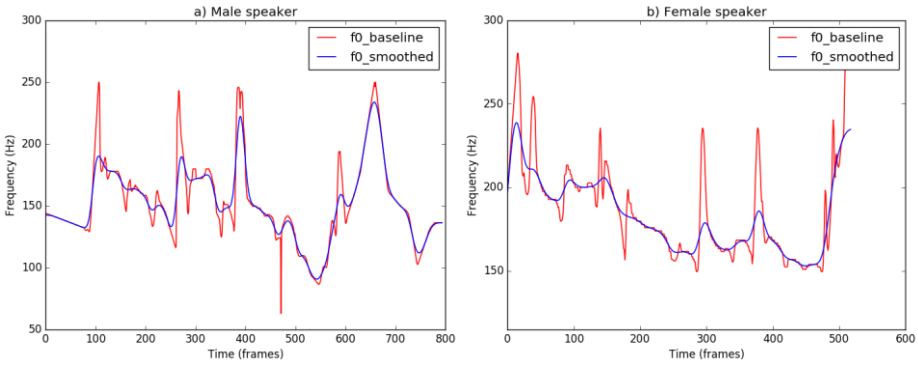


**Fig. 1.** Example of continuous F0 contour estimated by the baseline (red) and the smoothed approach (blue). Sentence: "Sometimes her dreams were filled with visions.", from both speakers.

### 3.2 Sinusoidal synthesis

Similarly to [10] [19], the synthesis procedure in this work decomposes the speech frames into a harmonic/voiced $s_v(t)$ component lower band and a stochastic/noise $s_n(t)$ component upper band based on MVF values. Thus, we define these components as

$$s(t) = s_v(t) + s_n(t) \tag{1}$$

The synthetic signal of the harmonic part is reconstructed by an OLA technique after generating the samples (harmonic amplitudes and phases) per each frame from their corresponding parameters. This can be expressed as:

$$s_v^i(t) = \sum_{k=1}^{K^i} A_k^i(\mathrm{t}) \cos\left(w_k^i t + \emptyset_k^i(t)\right) \tag{2}$$

$$w_k^i = 2\pi k(contF0)^i \tag{3}$$

where $A_k(\mathrm{t})$ and $\emptyset_k(t)$ are the amplitude and phase at frame $i$, $t = 0, 1, \ldots, N$ and $N$ is the length of the synthesis frame. $K$ is the time-varying number of harmonics that depends on the contF0 and MVF:

$$K^i = \begin{cases} round\left(\dfrac{MVF^i}{contF0^i}\right) - 1, & voiced\ frames \\ 0, & unvoiced\ frames \end{cases}$$

Next, the amplitudes of the harmonics are calculated by sampling the cepstral envelope, whereas their phases are obtained through a minimum-phase approach given by the cepstral envelope (for more mathematical details, see [19]). If the current frame is voiced, the synthesized noise part $n(t)$ is filtered by a high-pass filter $f_h(t)$ with cutoff frequency equal to the local MVF, and then modulated by its time-domain envelope $e(t)$ as detailed in our previous work [8]

$$s_n^i(t) = e^i(t)\left[f_h^i(t) * n^i(t)\right]$$

In this case, the phases are given random values, and the noise amplitude is obtained in the same way as the harmonic part. For unvoiced frames, the harmonic part is obviously zero and the synthetic frame is typically equal to the generated noise while the $f_h(t)$ step is omitted. The overall architecture is depicted in the block diagram as shown in Fig. 2. Hence, we refer to this vocoder as a continuous sinusoidal model (CSM).

## 4 Experiments

The speech data used in this study consist of a database recorded for the purpose of developing TTS synthesis. Two English speakers were chosen from the CMU-ARCTIC database [20], denoted AWB (Scottish English, male) and SLT (American English, female), which respectively consists of 1138 and 1132 sentences. The waveform sampling rate of the database is 16 kHz. In the vocoding experiments, 100 sentences (50

sentences from each speaker) were chosen randomly to be analyzed and synthesized with the baseline [8], proposed vocoder, WORLD[3] Vocoder [21] that is a fast and high-quality vocoder, and the TANDEM-STRAIGHT[4] vocoder [22] that has mostly become the state-of-the-art model in SPSS. The frame shift was set to 5 ms, while all other parameters remain at their default values.

In order to achieve our goals and to verify the effectiveness of the proposed method, objective and subjective evaluations were carried out in the next subsections.

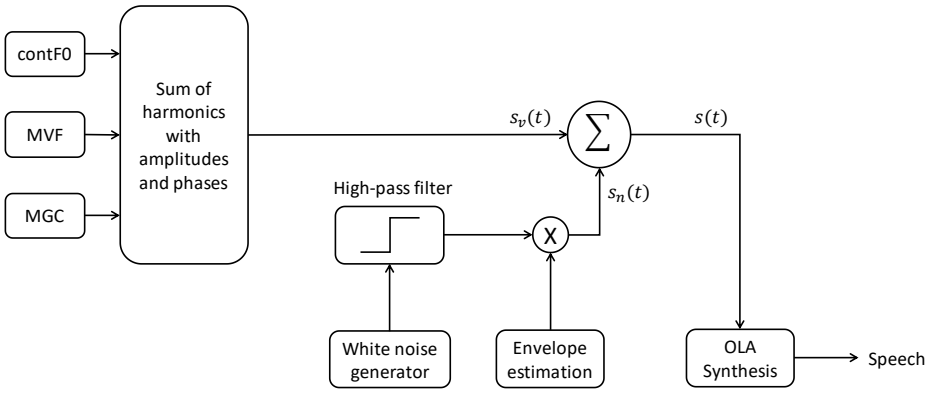


**Fig. 2.** Block diagram of the sinusoidal-synthesis part in a continuous vocoder.

### 4.1 Objective evaluation

A range of acoustic objective measures are considered to evaluate the quality of synthesized speech based on the proposed sinusoidal vocoder. We adopt the frequency-weighted segmental SNR ($fwSNR_{seg}$) for the error criterion since it is said to be much more correlated to subjective speech quality than classical SNR [23]. Moreover, Jensen and Taal [24] introduced an Extended Short-Time Objective Intelligibility (ESTOI) measure that calculates the correlation between the temporal envelopes of clean and processed speech. We also measure the Itakura-Saito (IS) distance that has played a key role in speech analysis and synthesis [25]. To a large extent, most studies (such as [26]) confirmed that when the IS distance is below 0.1, the two spectra would be perceptually nearly identical. For all objective measures, a calculation is done frame-by-frame and a higher value indicates better performance except for the IS measure (lower value is better). The results were averaged over the selected utterances (50 sentences) for each speaker.

As Table 1 shows, the proposed vocoder tends to significantly outperform the baseline approach among all metrics. In particular, it can be seen from IS measure that the proposed vocoder is slightly better than TANDEM-STRAIGHT in the AWB speaker whereas this is not the case with the SLT speaker. Hence, it can be concluded that the

[3] https://github.com/mmorise/World

[4] http://www.wakayama-u.ac.jp/~kawahara/tSTRAIGHT/TANDEM-STRAIGHTwithGUI.html

improved continuous vocoder presented in this paper has similar, or only slightly worse, performance to the reference vocoders.

**Table 1.** Average scores performance based on synthesized speech for male and female speakers. The bold font shows the best performance.

| Vocoder | IS | | fwSNRseg | | ESTOI | |
|---|---|---|---|---|---|---|
| | AWB | SLT | AWB | SLT | AWB | SLT |
| Baseline | 0.148 | 0.447 | 6.987 | 7.940 | 0.517 | 0.676 |
| Proposed | **0.058** | **0.082** | **9.560** | **11.034** | **0.749** | **0.867** |
| WORLD | **0.016** | **0.014** | **13.312** | 13.336 | **0.808** | **0.951** |
| TANDEM-STRAIGHT | 0.065 | 0.042 | 11.840 | **14.641** | 0.772 | 0.933 |

In addition, Table 2 compares the parameters of the vocoders under study. It can be seen that the continuous vocoder uses only two one-dimensional parameters for modeling the excitation, the WORLD vocoder is applying a five-dimensional band aperiodicity, whereas TANDEM-STRAIGHT use high-dimensional parameters which is not suitable for statistical modeling. As a result, continuous vocoder has fewer parameters; thus, its complexity of a synthesis algorithm is lower than those given in Table 2.

**Table 2.** Parameters and excitation type of applied vocoders

| Vocoder | Parameter per frame | Excitation |
|---|---|---|
| Continuous | F0: 1 + MVF: 1 + MGC: 24 | Mixed |
| WORLD | F0: 1 + Band aperiodicity: 5 + MGC: 60 | Mixed |
| TANDEM-STRAIGHT | F0: 1 + Aperiodicity: 2048 + Spectrum: 2048 | Mixed |

### 4.2 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 [27]. We compared natural sentences with the synthesized sentences from the baseline, proposed, TANDEM-STRAIGHT, WORLD, and an anchor system. The anchor type was the re-synthesis of the sentences with a standard pulse-noise excitation vocoder. 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[5].

13 participants (7 males, 6 females) with an age range of 20-42 years (mean: 31 years), mostly with engineering background were asked to conduct the online listening test. We evaluated ten sentences (five from each speaker). Altogether, 60 utterances were included in the test (2 speaker x 6 types x 5 sentences). On average, the test took

[5] http://smartlab.tmit.bme.hu/specom2018

15 minutes to fill. The MUSHRA scores for all the systems are showed in Fig. 3. According to the results, the proposed vocoder outperformed the baseline system for both speakers, but this difference is not statistically significant (Mann-Whitney-Wilcoxon ranksum test, $p<0.05$). For both speakers, the WORLD vocoder was ranked the best (for speaker AWB, this difference is significant). However, the proposed system was slightly preferred over TANDEM-STRAIGHT (not significant), showing that the sinusoidal extension of our vocoder is similar to state-of-the-art high quality vocoders. After checking the samples, we found that in case of speaker SLT, some of the synthesized samples using STRAIGHT contained clipping, which can explain the relatively weak performance of this vocoder

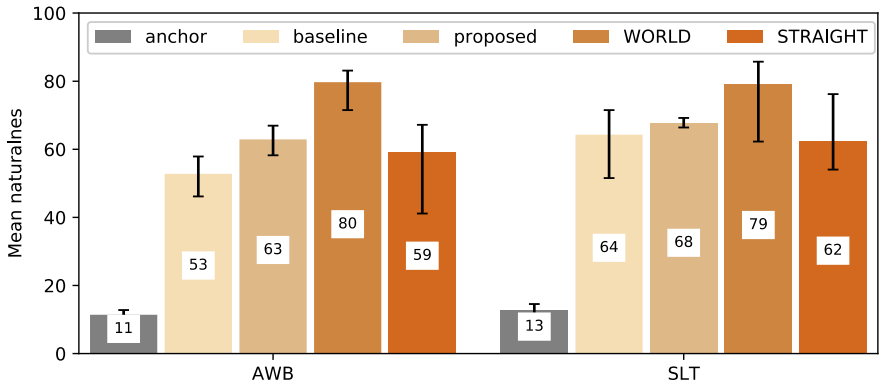


**Fig. 3** Results of the MUSHRA listening test for the naturalness question. Error bars show the bootstrapped 95% confidence intervals. The score for the reference (natural speech) is not included.

## 5 Conclusion

This paper proposed a new approach with the aim of designing a high quality continuous vocoder using a sinusoidal model. The performance of the systems has been evaluated through objective and subjective listening tests. Experiments demonstrate that our proposed model generates higher output speech quality than the baseline, that is a source-filter based model. Hence, our hypothesis was verified by successfully reconstructing the waveform from the same parameters used in the baseline. Moreover, it was found that the results obtained with the proposed vocoder were preferred over TANDEM-STRAIGHT and somewhat worse than with WORLD vocoders. The findings also point out that the continuous vocoder has few parameters and is computationally feasible; therefore, it is suitable for real-time operation.

For future work, the authors plan to train and evaluate all continuous parameters (F0, MVF, and MGC) using deep learning algorithm such as feed-forward and recurrent neural networks to test the proposed vocoder in statistical parametric speech synthesis.

**Acknowledgements.** The research was partly supported by the VUK (AAL-2014-1-183), by the EUREKA (DANSPLAT E!9944) projects. The Titan X GPU used for this

research was donated by NVIDIA Corporation. We would like to thank the subjects for participating in the listening test.

## References

[1] Zen, H., Tokuda, K., and Black, A., "Statistical parameteric speech synthesis," *Speech Communication,* vol. 51, no. 11, pp. 1039-1064, 2009.

[2] Kawahara, H., Masuda-Katsuse, I., and de Cheveign, A., "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication,* vol. 27, no. 3, p. 187–207, 1999.

[3] Drugman, T., Wilfart, G., Dutoit, T., "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," *In Proc. of Interspeech,* pp. 1779-1782, 2009.

[4] Morise, M., Yokomori, F., and Ozawa, K., "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems,* vol. 7, no. E99-D, pp. 1877-1884, 2016.

[5] Tamás Gábor Csapó, Géza Németh, and Cernak, M., "Residual-Based Excitation with Continuous F0 Modeling in HMM-Based Speech Synthesis," *3rd International Conference on SLSP 2015,* vol. 9449, pp. 27-38, 2015.

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

[7] Drugman, T., and Stylianou, Y., "Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra," *IEEE Signal Processing Letters,* vol. 21, no. 10, p. pp. 1230–1234, 2014.

[8] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh, "Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis," *Interspeech,* vol. Stockholm, no. , pp. 434-438, 2017.

[9] McAulay, R. J., and Quatieri, T. F., "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on ASSP,* vol. 34, no. 4, p. 744–754, 1986.

[10] Stylianou, Y., Laroche, J., and Moulines, E., "High-quality speech modification based on a harmonic + noise model," *in Proc. of Eurospeech,* pp. 451-454, 1995.

[11] Degottex, G., and Stylianou, Y., "A Full-Band Adaptive Harmonic Representation of Speech," in *In Proc. of Interspeech*, Portland, USA, 2012.

[12] Hu, Q., Stylianou, Y., Maia, R., Richmond, K., and Yamagishi, J., "Methods for applying dynamic sinusoidal models to statistical parametric speech synthesis," *IEEE ICASSP,* vol. South Brisbane, no. QLD, pp. 4889-4893, 2015.

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

[14] Drugman T. and Stylianou Y., "Maximum Voiced Frequency Estimation : Exploiting Amplitude and Phase Spectra," *IEEE Signal Processing Letters,* vol. 21, no. 10, p. pp. 1230–1234, 2014.

[15] Tokuda K., Kobayashi T. ,Masuko T., and Imai S., "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," *in Proc. of the ICSLP,* p. 1043–1046, 1994.

[16] Imai S., Sumita K., and Furuichi C., "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.

[17] Talkin, D., "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis,* pp. 497-518, 1995.

[18] Rabiner, L., Sambur, M., and Schmidt, C., "Applications of a nonlinear smoothing algorithm to speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 23, no. 6, pp. 552-557, 1975.

[19] Erro, D., Sainz, I., Navas, E., and Hernaez, I., "Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis," *IEEE Journal of Selected Topics in Signal Processing,* vol. 8, no. 2, pp. 184-194, 2014.

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

[21] Morise, M., Yokomori, F., and Ozawa, K., "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems,* vol. 7, no. E99-D, pp. 1877-1884, 2016.

[22] Kawahara, H., Morise, M., "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *Sadhana,* vol. 36, no. 5, pp. 713-727, 2011.

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

[24] Jensen J. and Taal C. H., "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, no. 11, pp. 2009-2022, 2016.

[25] Itakura, F., Saito, S., "An analysis-synthesis telephony based on the maximum-likelihood method," *Proc. Int. Congr. Acoust.,* vol. Japan, p. C17–C20, 1968.

[26] Chen, J., Benesty, J., Huang, Y., Doclo, S., "New insights into the noise reduction Wiener filter," *IEEE Audio Speech Lang. Process.,* vol. 14, no. 4, pp. 1218-1234, 2006.

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