Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Telecommunications and Media Informatics

High-Quality Vocoding Design with Signal Processing for Speech Synthesis and Voice Conversion

Ph.D. thesis booklet
Doctoral School of Informatics

Mohammed Salah Hamza Al-Radhi
M.Sc. in Communication Systems Engineering

Supervisors:
Tamás Gábor Csapó, Ph.D.
Géza Németh, Habil, Ph.D.

Budapest, Hungary
2020

1. Introduction

The quote "*All we need to do is make sure we keep talking*", said by Stephen Hawking, is to draw the encouraging line of all the speech technology. With the fast growth of computer technology to become more functional and prevalent, a wide range of the speech processing area is becoming a core function for establishing a human-computer communication interface. An excellent example of this, among other multimedia applications, is known both as the *speech synthesis*, i.e. artificial generation of speech waveforms, and as *text-to-speech*, i.e. building natural-sounding synthetic voices from text. Together are of great current interest and are still receiving much attention from researchers and industry.

State-of-the-art text-to-speech (TTS) synthesis is either based on unit selection or statistical parametric methods. In the last decade, particular attention has been paid to hidden Markov-model (HMM) based on TTS, which has gained much popularity due to its advantages in flexibility, smoothness, and small footprint. Deep neural networks (DNNs) have also become the most common types of acoustic models used in TTS for obtaining high-level data representations and the availability of multi-task learning, that a significant improvement in speech quality can be achieved. In addition, recent work has demonstrated that a WaveNet can generate close to human-level speech in TTS synthesis. In view of these systems, the speech signal is decomposed to parameters representing excitation (e.g. fundamental frequency, F0) and spectrum of speech, these are fed to a machine learning system. After the statistical model is learnt on the training data, during synthesis, the parameter sequences are converted back to speech signal with reconstructing methods (e.g. speech vocoders, excitation models).

Although nowadays TTS systems are intelligible, a limitation of current parametric techniques does not allow full naturalness yet and there is room for improvement in being close to human speech. According to recent summaries [1], there are three main factors in statistical parametric speech synthesis that are needed to deal with in order to achieve as high quality synthesized speech as unit selection: improved vocoder techniques, acoustic modeling accuracy, and over-smoothing during parameter generation. A number of such vocoders, also called as excitation models, were proposed in the last several years. According to a recent study, simple and uniform vocoders, which would handle all speech sounds and voice qualities in a unified way, are still missing. Although there are vocoding methods which yield in close to natural synthesized speech (e.g. the current de facto method, STRAIGHT [2]), they are typically computationally expensive, and are thus not suitable for real-time implementation, especially in embedded environments. Therefore, my *first hypothesis* is that: there is a need for simple and computationally feasible digital signal processing algorithms for generating high-quality and natural-sounding synthesized speech.

The goal of another related field of speech technology, *Voice Conversion* (VC), is to modify the speech of a source speaker with digital speech signal processing methods in a way that it will be similar to the speech of a target speaker, while the linguistic content remains the same. Although there has been a long research in voice conversion [3], there is still space for improvement as current methods lack the flexibility to convert the speech of any source speaker to any other target speaker. Moreover, the naturalness of the converted voice still deteriorates compared to the source speaker due to over-smooth phenomenon or discontinuity problem which makes the converted speech sound muffled. Therefore, improving the performance of converted voice is still a challenging research question. My *second hypothesis* is that: there is a need to develop advanced adaptable vocoder based VC for achieving high-quality converted speech.

Hence, this dissertation proposes a solution to achieve higher sound quality and conversion accuracy with the deep learning advances, while its approach remains computationally efficient.

2. Research objectives

The main goal of my Thesis work is to construct a vocoder that is very flexible (whose parameters can be controlled) with respect to achieving high quality synthesized speech. This challenge required five major contributions of the work presented in this thesis booklet, as depicted in Figure 1:

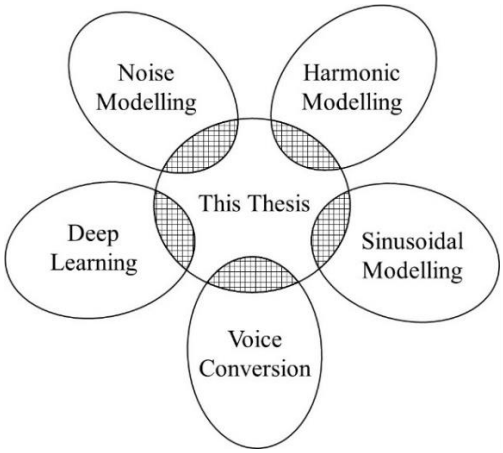


Figure 1. Thesis contribution

The *first research* objective is **modulating the noise component of the excitation signal**. It was argued that the noise component is not accurately modelled in modern vocoders (even in the widely used STRAIGHT vocoder). Therefore, two new techniques for modelling voiced and unvoiced sounds are proposed in this part of the research by: 1) estimating the temporal envelope of the residual signal that is helpful in achieving accurate approximations compared to natural speech, and 2) noise masking to reduce the perceptual effect of the residual noise and allowing a proper reconstruction of noise characteristics.

The *second research* objective is **harmonic modelling**. This study focuses on improving the accuracy of the continuous fundamental frequency estimation and the naturalness of the speech signal by proposing three different adaptive techniques based on Kalman filter, time warping, and instantaneous frequency. A clear advantage of the proposed approaches is its robustness to additive noise. Moreover, Harmonic-to-Noise ratio technique is added as a new vocoded-parameter to the voiced and unvoiced excitation signal in order to reduce the buzziness caused by the vocoder.

The *third research* objective is **acoustic modelling design based on deep learning**. In this part of the research, the novel continuous vocoder was applied in the acoustic model of deep learning based speech synthesis using feedforward and recurrent neural networks as an alternative to HMMs. Here, the objective is two-fold: (a) to overcome the limitation of HMM which often generate over-smoothing and muffled synthesized speech, (b) to ensure that all parameters used by the proposed vocoder were taken through training that could synthesize very high quality TTS.

The *fourth research* objective is **designing a sinusoidal modelling system**. Here, a new continuous vocoder was built using a sinusoidal model that is applicable in statistical frameworks

which decomposes the speech frames into a harmonic component lower band and a stochastic component upper band based on maximum voiced frequency. The objective is to consider whether a different synthesis technique can produce more accurate synthesis of speech than the source-filter model.

The *fifth research* objective is **proposing a novel model applied for voice conversion with parallel training data**. This part of research includes investigating the novel continuous vocoders in a VC framework. The vocoders are tested both in same-gender and cross-gender scenario.

As a final point, this Thesis booklet provides a detailed and complete study on several speech analysis and synthesis techniques and their applications in text-to-speech and voice conversion.

3. Methodology

I validated the proposed research by experiments and analytical examinations, in which I developed and improved a novel continuous vocoder for SPSS. The applied methodology of this dissertation follows the international standards. In the following, speech databases, conditions, and evaluation methods are detailed.

3.1. Continuous vocoder: Baseline

The first version of a residual-based vocoder was proposed by [4], using a continuous F0 (contF0) [5], and maximum voiced frequency (MVF) [6]. 24-order Mel-generalized cepstral analysis (MGC) [7] 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 MGC parameter streams.

During the synthesis phase, voiced excitation is composed of principle component analysis (PCA) residuals overlap-added pitch synchronously, depending on the contF0. 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. Voiced and unvoiced excitation is added together. Finally, a Mel generalized-log spectrum approximation (MGLSA) filter is used to synthesize speech from the excitation and the MGC parameter stream [8].

3.2. Speech corpora

In order to evaluate the performance of the suggested models, a database containing a few hours of speech from several speakers was required for giving indicative results. Five English speakers were firstly chosen from the CMU-ARCTIC[1] database [9], denoted BDL (American English, male), JMK (Canadian English, male), AWB (Scottish English, male), CLB (US English, female), and SLT (American 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% triphone coverage. The waveform sampling rate of this database is 16 kHz.

[1] http://www.festvox.org/cmu_arctic/

My second database is the corpus created by my co-author in [J1]. It was motivated by the fact that it builds the first modern standard Arabic audio-visual expressive corpus which is annotated both visually and phonetically. It contains 500 sentences with 6 emotions (Happiness – Sadness – Fear – Anger – Inquiry – Neutral), and recorded by a native Arabic male speaker (denoted ARB). The waveform sampling rate of this database is 48 kHz.

The third corpus is the one based on Hungarian language. Two Hungarian male and two female subjects (4 speakers) with normal speaking abilities were recorded while reading sentences aloud (altogether 209 sentences each). The ultrasound data and the audio signals were synchronized using the tools provided by Articulate assistant advanced software [C6]. The waveform sampling rate of this database is 44 kHz.

3.3. Performance Measurement and Metrics

A range of objective speech quality and intelligibility measures are considered to evaluate the quality of synthesized speech based on the proposed methods. I adopt the frequency-weighted segmental SNR (fwSNRseg), extended short-time objective intelligibility (ESTOI), weighted-slope spectral distance (WSS), normalized covariance metric (NCM), and mel-cepstral distortion measures [10] [11] [12]. Also, I used a series of distinct measurements in accordance with [13] to assess the accuracy of my third-Thesis as discussed in Section 4. These are gross pitch errors (GPE), mean fine pitch errors (MFPE), and standard deviation (SD). Additionally, I compared the natural and vocoded sentences by measuring the phase distortion deviation (PDD) as it carries all of the crucial information relevant to the shape of glottal pulses [14]. For all these objective measures, a calculation is done frame-by-frame and the results were averaged over the tested utterances for each speaker.

3.4. Reference System

The proposed vocoder based on TTS and VC systems was evaluated by comparing it with strong reference systems. STRAIGHT [2] and WORLD [15] are high-quality vocoders and widely regarded the state-of-the-art model in SPSS. MagPhase [16] and log domain pulse model (PML) [17] are considered as a modern sinusoidal models. Sprocket [18] is a vocoder-free VC system that was used recently for the voice conversion challenge 2018. YANGsaf algorithm [19] is an F0 estimator method that can be compared along with adaptive contF0 in Thesis 3. The choice of YANGsaf is confirmed by the fact that it was recently shown in [20] to outperform other well-known F0 estimation approaches found in the current literature (like YIN, RAPT, or DIO).

3.5. Experimental conditions

I used the open source Merlin[2] [21] to implement machine learning methods introduced in Thesis III and Thesis V, constructive changes are also introduced in this toolkit to be able to adapt the proposed vocoder. The training sets contain 90% of the speech corpora, while the rest were used for testing. The training procedures were conducted on a high-performance NVidia Titan X GPU. In the vocoding experiments, 100 sentences from each speaker were analyzed and synthesized with the baselines and proposed vocoders.

[2] https://github.com/CSTR-Edinburgh/merlin

With the purpose of assessing true performance of the continuous pitch tracking presented in Thesis II, a reference pitch contour (ground truth) is estimated from the electro-glottalgraph (EGG) as it is directly derived from glottal vibration and is largely unaffected by the nonharmonic components of speech. In my evaluation, the ground truth is extracted from EGG signals using Praat [22]. The analysis of the measurements and the statistical examinations were made by Python 3.5 and MATLAB 2018b.

3.6. Perceptual listening test

I conducted several web-based MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) listening tests [23] in order to evaluate which system is closer to the natural speech. I compared natural sentences with the synthesized sentences from the baseline, proposed, and a hidden anchor system (different for each test). Listeners were asked before the test to listen to an example from a speaker to adjust the volume. 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 (different for each participant).

Besides, Mean Opinion Score (MOS) test was also carried out in Chapter 10 and 11. In the MOS test, I compared three variants of the sentences: 1) Target, 2) Converted speech using the baseline systems, and 3) Converted speech using the proposed vocoder. Similarly to the MUSHRA test, the listeners had to rate the naturalness of each stimulus, from 0 (highly unnatural) to 100 (highly natural).

About 20 participants (for each test) between the age of 23-40 (mean age: 30 years), mostly with engineering background, were asked to conduct the online listening tests. Altogether, 10 MUSHRA with 2 MOS tests were performed during my research work to evaluate my dissertation. On average, the MUSHRA test took 14 minutes, while the MOS test was 12 minutes long.

4. New scientific results

The following subsections describe the contribution of my Thesis booklet. The scientific results of all following theses are submitted and published in international journals and conferences.

4.1. Thesis group I: Modulating the Noise Component of Excitation Signal

Since the design of a vocoder-based SPSS depends on speech characteristics, the preservation of voice quality in the analysis/synthesis phase and the irregular “buzzy” synthetic speech sounds are the main problems of the vocoder. [24] presents an experimental comparison of a wide range of important vocoder types which have been previously invented. Despite the fact that most of these vocoders have been successful in synthesizing speech, they are not successful in synthesizing high-quality and natural-sounding speech. The reason for this is the inaccurate composition and estimation of the vocoder parameters which leads to a degradation in the speech signal.

Therefore, this Thesis group considers the above issues by suggesting robust methods for advanced modeling of the noise excitation which can yield an accurate noise component of the excitation to remove the buzzy quality, while the vocoder remains still computationally efficient. Accordingly, Figure 2 shows the main components of the proposed continuous vocoder.

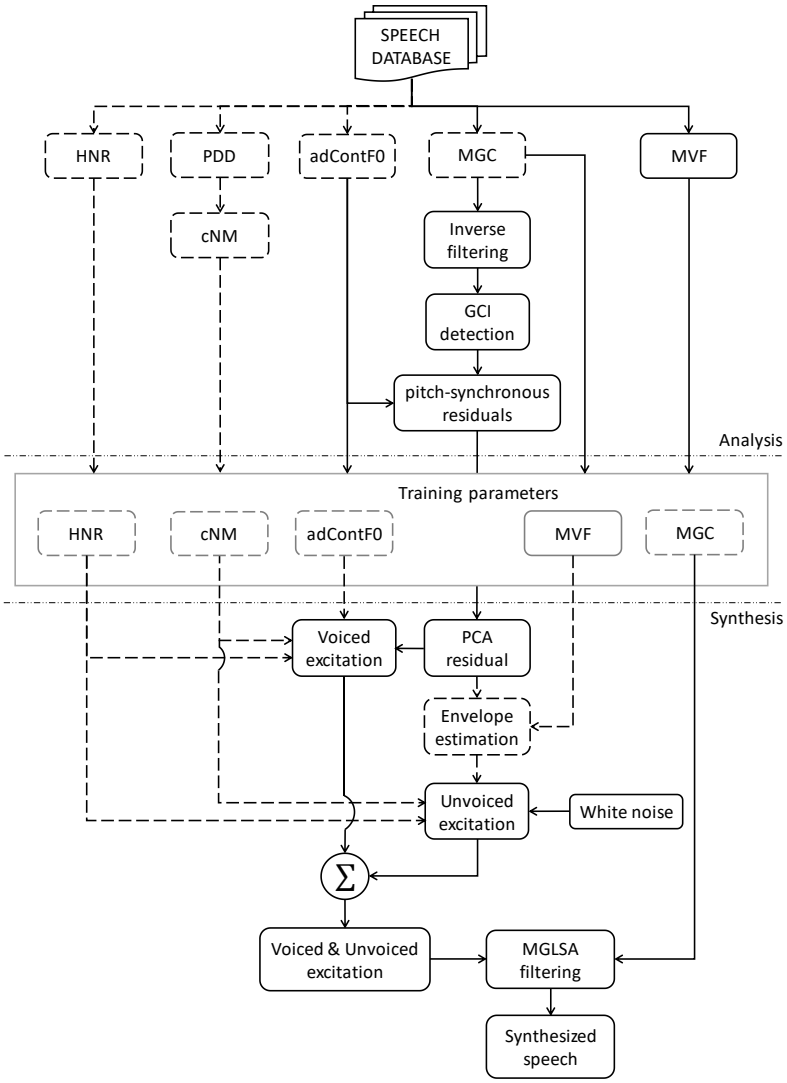


Figure 2. Schematic diagram of the developed continuous vocoder. Additions and refines are marked with dashed lines.

Thesis I.1: Temporal envelopes [C1, J1, J5] *I designed and implemented a new method to shape the high-frequency component of the unvoiced excitation by estimating the time envelope of the residual signal. I showed that this approach is helpful in achieving accurate approximations compared to natural speech and produce synthetic voice with significantly better quality than the baseline.*

In the baseline (Section 3.1), there is a lack of voiced component in higher frequencies. However, it was shown that in natural speech, the high-frequency noise component is time-aligned with the pitch (F0) periods. Therefore, I designed a temporal envelope to shape the high-frequency component (above MVF of the excitation) by estimating the envelope of the PCA residual and modifying the noise component by this envelope to make it more similar to the residual of natural speech. Amplitude, Hilbert, Triangular, and True envelopes are investigated, enhanced, and then

applied to the noise component of the excitation in the continuous vocoder to present a more reliable envelope. I have also proposed that the True envelope with weighting factor will bring a unique time envelope which makes the convergence more closely to the natural speech. The natural and vocoded sentences were compared by measuring the PDD. It was found that the baseline vocoding sample has too much noise component compared to the natural sample. On the other hand, the proposed systems with envelopes have PDD values closer to the natural speech. In particular, the 'Amplitude' envelope system results in too low PDD values, meaning that the noisiness is too low compared to natural speech. Otherwise, Hilbert envelope brings out some hidden information more efficiently and fits a curve that approximately matches the peaks of the residual frame as shown in Figure 3.

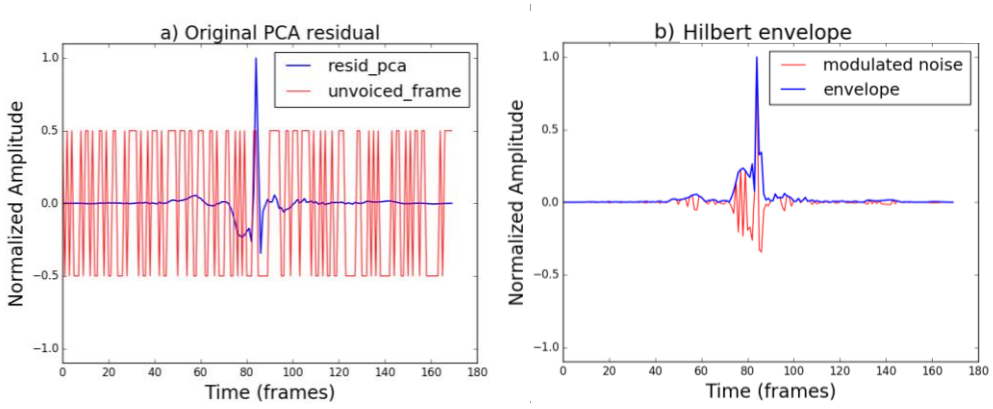
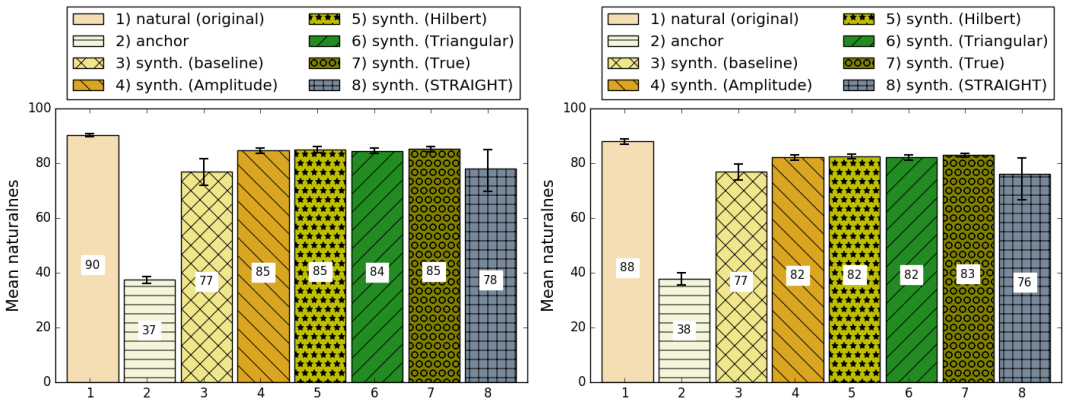


Figure 3. Illustration of the performance of the time envelope. "unvoiced_frame" is the excitation signal consisting of white noise, whereas "resid_pca" is the result of applying PCA on the voiced excitation frames.

The MUSHRA scores of the listening test are presented in Figure 4. The listening test samples can be found online[3]. It can be observed that all of the proposed systems significantly outperformed the baseline (Mann-Whitney-Wilcoxon rank sum test). Moreover, a significant improvement was noted in sound quality with the proposed systems over STRAIGHT vocoder. I therefore draw the conclusion that the approach presented in this Thesis is an interesting alternative to the earlier version of the continuous vocoder, and reached higher naturalness scores in the listening test than those of STRAIGHT vocoder.

[3] http://smartlab.tmit.bme.hu/vocoder_Arabic_2018



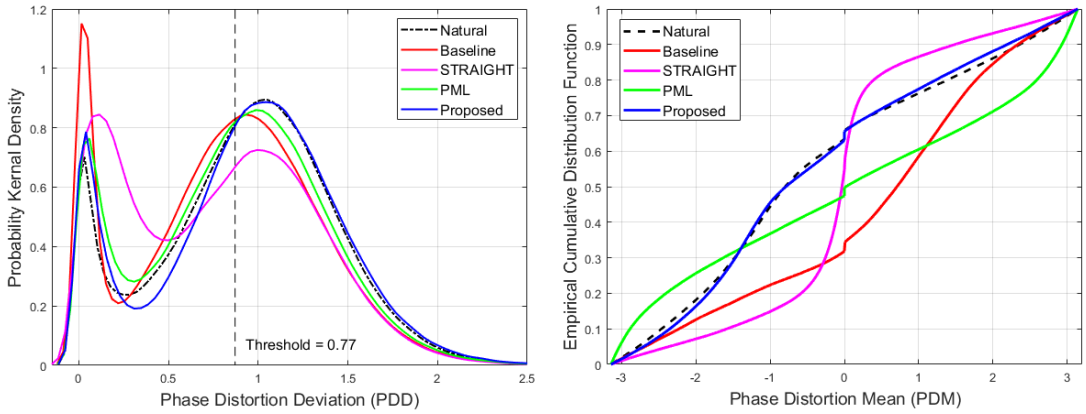
a) Arabic samples b) English samples

Figure 4. Results of the subjective evaluation. Higher value means larger naturalness. Error bars show the bootstrapped 95% confidence intervals.

Thesis I.2: Continuous Noise Masking [J4] *I proposed an algorithm based on noise masking to reduce the perceptual effect of the residual noise and allowing a proper reconstruction of noise characteristics. I proved experimentally that continuous noise masking gives better quality resynthesized speech than traditional binary masking techniques*.

Traditional parametric vocoders generally show a perceptible deterioration in the quality of the synthesized speech due to different processing algorithms. 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, I proposed in this Thesis a new masking approach called continuous NM that changes from 0 to 1 (or 1 to 0) rather than a binary 0 or 1 as in the conventional binary NM. In other words, if the value of the continuous NM 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). This means that cNM can save parts of speech component in the weak-voiced and unvoiced segments. This offers a flexibility to represent voiced and unvoiced segments in a uniform way under a condition of the continuous NM threshold.

In terms of experimental results, density estimate using a kernel smoothing method 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 (Figure 5a), and empirical cumulative distribution function of phase distortion mean values are also calculated in Figure 5b to see whether these systems can be normally distributed and how far they are from the natural signal. It could be clearly observed that the proposed method significantly outperforms all baseline vocoders.



(a) (b)

Figure 5. a) Estimation of the probability kernel density functions of PDDs, and b) Empirical cumulative distribution function of PDMs.

In order to evaluate the perceptual quality of the proposed system, we conducted a web-based MUSHRA listening test. The listening test samples can be found online[4]. Based on the overall results, I can conclude that among the techniques investigated in this Thesis of noise reconstruction, cNM performs well in continuous vocoder when compared with other approaches as shown in Figure 6. In other words, the difference between STRAIGHT, PML and the proposed vocoders is not statistically significant (Mann-Whitney-Wilcoxon ranksum test, $p<0.05$), meaning that the proposed vocoder reached the quality of state-of-the-art results.

I therefore draw the conclusion that the approach reported in this Thesis is beneficial and can give either accurate noise reconstruction for some voices that involve the presence of noise in voiced segments (e.g., breathy voice), or reduce any other residual buzziness. In particular, the cNM parameter is not limited only to our vocoder, it can be also applied it to other types of modern parametric vocoders (such as Ahocoder [25] or even the PML vocoders) for better synthesis of the noise in voiced segments.

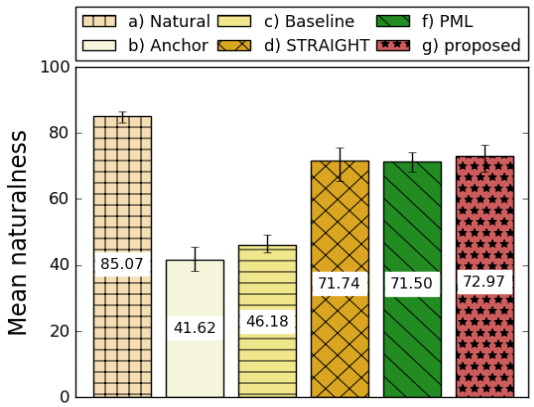


Figure 6. Results of the subjective evaluation for the naturalness question. A higher value means larger naturalness. Error bars show the bootstrapped 95% confidence intervals.

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

4.2. Thesis group II: Harmonic Modeling

This Thesis group concerns with estimating the fundamental frequency on clean and noisy speech signals. In particular, continuous F0 is still sensitive to additive noise in speech signals and suffers from short-term errors (when it changes rather quickly over time). Moreover, contF0 can cause some tracking errors when the speech signal amplitude is low or the voice is creaky. Therefore, I described novel approaches which can be used to enhance and optimize some other existing F0 estimator algorithms. Additionally, the Harmonic-to-Noise ratio technique is added as a new vocoded-parameter to the voiced and unvoiced excitation signal in order to reduce the buzziness caused by the vocoder.

Recent studies in TTS synthesis have shown the benefit of using a continuous pitch estimate; one that interpolates fundamental frequency even when voicing is not present. But, contF0 still sensitives to additive noise in speech signals and suffers from short-term errors (when it changes rather quickly over time). Moreover, contF0 can cause some tracking errors when speech signal amplitude is low, voice is creaky, or low HNR. To alleviate these issues, three adaptive techniques have been developed in this Thesis for achieving a robust and accurate contF0: 1) we weight the pitch estimates with state noise covariance using adaptive Kalman-filter framework, 2) we iteratively apply a time axis warping on the input frame signal, 3) we optimize all F0 candidates by using instantaneous-frequency based approach.

The accuracy of the proposed algorithms is validated using a database of human recordings where the glottal cycles are calculated from electroglottograph (EGG) signals. Results based on objective and perceptual tests demonstrate that these approaches achieve significantly higher accuracy and smoother contF0 trajectory on noisy and clean speech.

Thesis II.1: Adaptive Continuous Pitch Algorithm [C5, C10, J2] *I developed and applied an adaptive approach based on Kalman filtering, time warping, and instantaneous frequency to optimize the performance of the continuous F0 estimation algorithm in clean and noisy speech. I showed that a clear advantage of the proposed approach is its robustness to additive noise; and the voice built with the proposed framework gives state-of-the-art speech synthesis performance while outperforming the previous baseline.*

Adaptive Kalman Filtering: It is known from the literature that Kalman filter is the optimal state estimation method for a stochastic signal when the noise of the state and the noise of the measurements are Gaussian, and their covariance matrixes is expected to be known. However, this can be very difficult in practice. If the noise statistics (estimates of the state and measurement noise covariance) are not as expected, Kalman filter will be unstable or gives state estimates that are not close to the true state. One promising approach to overcome this problem is the use of adaptive mechanisms into a Kalman filter. In particular, signal quality indices (SQIs) have been proposed by [30], which give the confidence in the measurements of each source. When the SQI is low, the measurement should not be trusted; this can be achieved by increasing the noise covariance. In spite of this, the state noise was a priori fixed in [30].

Therefore, to improve the contF0 estimation method, I used the SQI algorithm with Kalman filter in order to compute the confidence in both state noise and measurement noise covariance. So that, their covariance matrices are updated appropriately at each time step until convergence. Table

1 displays the results of the evaluation of the proposed method based on contF0, for female and male speakers, in comparison to the YANGsaf algorithm [19]. Accordingly, the findings in Table 1 strongly supports the use of proposed method based on adaptive Kalman filter (AKF) as the most accurate contF0 estimation algorithm over the baseline method. Moreover, the proposed #1 vocoder based on AKF in Figure 7 clearly outperformed the baseline system in the listening test.

I can conclude that the main advantage of using adaptive Kalman filter is that I can determine our confidence in the estimates of contF0 algorithm based TTS by adjusting SQIs to update both the measurement noise covariance and the state noise covariance. For example, it can be used to replace the one studied by Li et al. [30] in the heart rate assessment application.

Adaptive Time-Warping: In the speech signal, it is necessary that harmonic components are separated from each other with the purpose of being easily found and extracted. Once F0 rapidly changes, harmonic components are subject to overlap each other and make it difficult to separate these components; or the close neighboring components make the separation through filtering very hard especially with a low spectral voice (such as male pitch) [31]. To overcome this problem, previous work in the literature has provided method by introducing a time-warping based approach.

To achieve a further reduction in the amount of contF0 trajectory deviation (deviate from their harmonic locations) and to avoid additional sideband components generation when a fast movement of higher frequencies occurs, adaptive time warping approach can be used to refine the contF0 algorithm. Firstly, stretching the time axis in order to make the observed contF0 value in the new temporal axis stay unchanged and preserves the harmonic structure intact. The second step is that the input waveform is filtered by bandpass filter bank with different center frequencies f_c to separate only the fundamental component in the range near f_c. weighted average of F0 information yields the F0 estimate on the warped time axis. Converting this estimate value to the value on the original time axis provides the further improved F0 estimate. Recursively applying these steps gives a final adaptive contF0 estimate

I test this new F0 tracker in two different ways: additive noise and online listening test. Accordingly, the results of white noise suggest that the error metrics for the proposed method based on adaptive time warping (TWRP) are smaller than the baseline (Table 1), and proposed #2 vocoder based TWRP clearly outperformed the baseline system in the listening test (Figure 7).

I can conclude that the main advantage of using adaptive time warping scheme is that it has the ability to track the time-varying contF0 period and reduce the amount of contF0 trajectory deviation from their harmonic locations.

Adaptive Instantaneous Frequency: Instantaneous frequency is defined as the derivative of the phase of the waveform. Flanagan's equation [32] is used to calculate the instantaneous frequency $IF(t)$. Actual refinement is carried out using recursively the following equation

$$adaptive\ contF0 = \frac{\sum_{k=1}^{k} |S(kw_0)|\ IF(kw_0)}{\sum_{k=1}^{k} k|S(kw_0)|} \tag{1}$$

where w_0 represents the angular frequency of the contF0 candidate, $S(w)$ is the spectrum of a waveform, and k represents the harmonic number (we set $k = 6$ for further refinement of the methodology). This approach (named as a StoneMask) is also used in WORLD, that is a high-quality speech analysis/synthesis system, to adjust its fundamental frequency named DIO algorithm.

The impact of this approach on contF0 performance is illustrated in Table 1 and Figure 7. It is quite obvious that there is no significant difference between contF0 based on adaptive StoneMask (STMSK) and the state-of-the-art YANGsaf approaches based on MFPE and STD measures in all speakers. In perceptual test, the difference between STRAIGHT and the proposed #3 system (Figure 7) is not statistically significant (Mann-Whitney-Wilcoxon ranksum test, $p<0.05$), meaning that our methods reached a point of high naturalness of synthesized speech.

I can conclude that by considering the system processing speed, adaptive contF0 based on instantaneous frequency is computationally inexpensive and can be useful in a practical speech processing application.

Table 1. Average scores performance per each speaker in the presence of additive white noise (SNR = 0 dB). Lower value indicates better performance.

Method	GPE %			MFPE			STD		
	BDL	**JMK**	**SLT**	**BDL**	**JMK**	**SLT**	**BDL**	**JMK**	**SLT**
baseline	33.170	40.057	27.502	4.050	3.901	3.512	4.393	4.293	3.912
contF0_AKF	31.728	40.865	26.122	3.211	3.241	2.898	3.465	3.627	3.448
contF0_TWRP	29.464	37.839	26.932	3.199	3.165	2.890	3.449	3.511	3.186
contF0_STMSK	31.418	37.052	26.352	2.128	1.896	2.067	2.103	1.658	2.058
YANGsaf	27.530	35.200	25.852	2.233	2.181	2.175	2.206	2.219	2.265

Thesis II.2: Parametric HNR Estimation Approach [C8, J2] *I proposed the addition of a new excitation HNR parameter to the voiced and unvoiced components. I proved that it can indicate the degree of voicing in the excitation and reduce the buzziness caused by the vocoder*.

The main goal of vocoders is to achieve high naturalness speech intelligibility. It was shown in Thesis I.1 that the temporal envelope yields sufficiently good quality in the synthesized speech by reducing the buzziness. Such an analysis/synthesis system may also suffer from some other degradations: 1) loss of the high-frequency harmonic components, 2) high-frequency noise components, or 3) noise components in the main formants. As the degree of these issues increases, more noise appears and consequently degrade the speech quality highly.

In this Thesis, I propose to add a continuous HNR as a new excitation parameter to the continuous vocoder in order to alleviate previous problems. The HNR is positive infinite for purely harmonic sounds while it is very low for the noise. In a continuous vocoder, my approach here is to use the HNR to weight the excitation signal in both voiced and unvoiced frames. As a result,

the voiced and unvoiced speech signal are added in the ratio suggested by the HNR, and then used to excite the MGLSA filter as illustrated in the bottom part of Figure 2.

The performance evaluation is summarized in Table 2. Focusing on the WSS, it is clear that proposed vocoder can outperform the STRAIGHT vocoder for JMK speaker. The NCM measure shows similar performance between proposed system and STRAIGHT. On the other hand, a good improvement was noted for the developed method in ESTOI measure. Thus, these experiments showing that adding HNR to the baseline vocoder was beneficial.

The MUSHRA scores of the listening test are presented in Figure 7. The listening test samples can be found online[5]. It can be observed that all of the proposed systems significantly outperformed the baseline (Mann-Whitney-Wilcoxon rank sum test). Moreover, a good improvement was noted in sound quality with the proposed #3 system over STRAIGHT vocoder. I therefore draw the conclusion that the experimental results demonstrated that the proposed methods can improve the naturalness of the synthesized speech over the earlier baseline and reached the state-of-the-art performance.

Table 2. Average scores performance based on synthesized speech signal per each speaker. Higher value indicates better performance except for the WSS.

Metric	Speaker	Baseline	STRAIGHT	Proposed
NCM	BDL	0.650	0.992	0.913
	JMK	0.620	0.963	0.906
	SLT	0.673	0.991	0.910
ESTOI	BDL	0.642	0.923	0.892
	JMK	0.620	0.895	0.873
	SLT	0.679	0.945	0.894
WSS	BDL	48.569	22.144	24.013
	JMK	51.788	29.748	26.238
	SLT	58.043	23.614	26.906

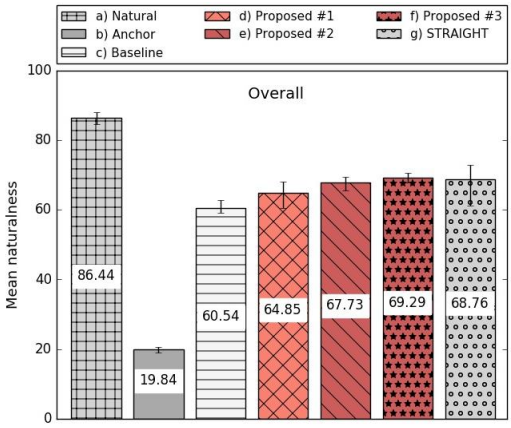


Figure 7. Results of the subjective evaluation for the naturalness question. A higher value means larger naturalness. Error bars show the bootstrapped 95% confidence intervals.

[5] http://smartlab.tmit.bme.hu/adContF0_2019

4.3. Thesis group III: Acoustic Modeling Based on Deep Learning

The popularity of hidden Markov models (HMMs) has been growing over the past decade, motivated by its accepted advantages of convenient statistical modelling and flexibility. Even though the quality of synthesized speech generated by HMM based speech synthesis has been improved recently, its naturalness is still far from that of actual human speech. Moreover, these models have their limitations in representing complex and nonlinear relationships between the speech generation inputs and the acoustic features. In the last few years, deep machine learning algorithms have shown in many domains their ability to extract high-level, complex abstractions, and data representations from large volumes of supervised and unsupervised data [29].

Therefore, the goal of the work reported in this Thesis was to apply a continuous vocoder in deep neural network based on TTS. The experiments were successful and I was able to add continuous features (contF0, MVF, and MGC) to the model of the deep learning for high-quality speech synthesis.

Thesis III.1: Feed-Forward Deep Neural Network [C2, C9, J5] *I built and implemented deep learning based acoustic modeling using FF-DNN with the continuous vocoder. The proposed DNN-TTS significantly outperformed the baseline method based on HMM-TTS, and its naturalness approaches the high-quality WORLD vocoder based TTS.*

The baseline system [27] was successfully used with HMM-based TTS. However, HMMs often generate over-smoothing, and muffled synthesized speech. Recently, neural approaches have achieved significant improvements to replace the decision tree used in HMM-based speech [28], and shown their ability to model high-dimensional acoustic parameters and the availability of multi-task learning [29]. From these points, we propose a training scheme for multilayered perceptron which tries to use the modified version of the continuous vocoder in DNN based TTS for further improving its quality.

The DNN-TTS used in this work is a feed-forward multilayered architecture with six layers of hidden units, each consisting of 1024 units. The input is used to predict the output with multiple layers of hidden units, each of which performs a non-linear function of the previous layer's representation, and a linear activation function was used at the output layer. Weights and biases were prepared with small nonzero values, and optimized with stochastic gradient descent to minimize the mean squared error between its predictions and acoustic features of the training set. Textual and phonetic parameters are first converted to a sequence of linguistic features as input, and neural networks are employed to predict acoustic features as output for synthesizing speech.

In order to evaluate the perceptual quality of the proposed system, Figure 8 shows the results of MUSHRA test based on DNN-TTS. The listening test samples can be found online[6]. Based on the overall results, the DNN-TTS with the continuous vocoder significantly outperformed baseline method based on HMM-TTS, and its naturalness is almost reached the quality of state-of-the-art WORLD vocoder based TTS. Hence, I can conclude that this Thesis showed the potential of the DNN-based approach for SPSS over the HMM-TTS.

[6] http://smartlab.tmit.bme.hu/vocoder2019

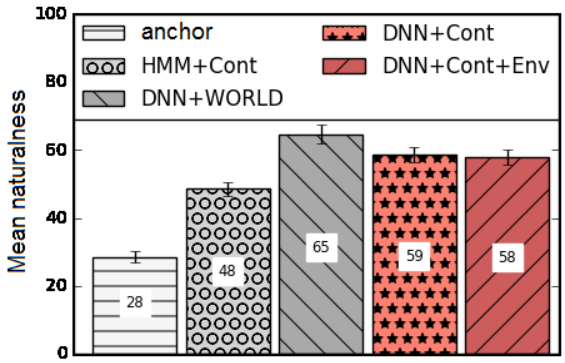


Figure 8. Scores of the MUSHRA listening test.

Thesis III.2: Sequence-to-Sequence Recurrent Neural Network [C3, C9, J5] *I investigated and examined sequence-to-sequence modelling using recurrent neural networks for the continuous vocoder. I showed that the performance of the vocoder can be significantly enhanced by the RNN framework and confirmed its superiority against the FF-DNN solution.*

In my recent work in SPSS, I developed a vocoder which was successfully used with a FF-DNN. However, I found some limitations of the conventional DNN-based acoustic modeling for speech synthesis, e.g. its lack of ability to predict variances, unimodal nature of its objective function, and the sequential nature of speech is ignored. In order to avoid these problems, I propose the use of sequence-to-sequence modeling with RNNs. The RNN consists of a set of stacked fully connected layers, where neurons can receive feedback from other neurons at the previous, same and next layer at earlier time steps.

Four neural network architectures (long short-term memory (LSTM), bidirectional LSTM (BLSTM), gated recurrent network (GRU), and Hybrid model based RNN with BLSTM) are investigated and applied using this continuous vocoder to model contF0, MVF, and MGC for more natural sounding speech synthesis. From both objective metrics (Table 3) and subjective evaluation[7][C3], experimental results demonstrated that the proposed RNN models can improve the naturalness of the speech synthesized significantly over the DNN baseline. In particular, the BLSTM network achieves better performance than others. Hence, I can conclude that the proposed continuous vocoder within the BLSTM framework performed well and reached the highest naturalness scores among other neural network topologies.

[7] http://smartlab.tmit.bme.hu/vocoder2019

Table 3. Objective measures for all training systems based on synthesized speech signal using continuous vocoder for SLT (female) and AWB (male) speakers. Lower value indicates better performance except for the CORR.

Systems	MCD (dB)		MVF (Hz)		F0 (Hz)		CORR		Validation error	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
DNN	4.923	4.592	0.044	0.046	17.569	22.792	0.727	0.803	1.543	1.652
LSTM	4.825	4.589	0.046	0.047	17.377	23.226	0.732	0.793	1.526	1.638
GRU	4.879	4.649	0.046	0.047	17.458	23.337	0.731	0.791	1.529	1.643
BLSTM	4.717	4.503	0.042	0.044	17.109	22.191	0.746	0.809	1.517	1.632
Hybrid	5.064	4.516	0.046	0.044	18.232	22.522	0.704	0.805	1.547	1.627

4.4. Thesis group IV: Sinusoidal Modeling

In recent years, a number of sophisticated source-filter based vocoders have been proposed and extensively used in speech synthesis. Sinusoidal vocoder is an alternative category for the source-filter model and has been successfully applied to a broad range of speech processing problems such as speech coding and modification [33]. 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.

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 [34] or intermediate parameters (regularized cepstral coefficients) [35] to overcome these issues, the computational complexity of SPSS can be quite high once additional algorithms and parameters are including [1].

Therefore, the goal of the work reported in this Thesis was to develop a new sinusoidal model as an alternative synthesis technique in a continuous vocoder, which can provide a high quality sinusoidal model with a fixed and low number of parameters. Listening test results show that in the same number of parameters, the suggested model is preferred to the STRAIGHT vocoder.

Thesis IV.1: Continuous Sinusoidal Model [C4] *I designed a new vocoder based on the sinusoidal model that is applicable in statistical frameworks. I validated the efficiency and quality of the proposed model and proved that the proposed CSM vocoder gives state-of-the-art performance in resynthesized speech while outperforming the source-filter vocoder.*

In [J4], we proposed a source-filter based vocoder which was successfully used with deep learning. Previous studies have shown that human voice can be modelled effectively as a sum of sinusoids. Therefore, I address the design of a continuous vocoder using sinusoidal synthesis model with a minimum phase that is applicable in statistical frameworks. The synthesis procedure described in this vocoder decomposes the speech frames into a harmonic/voiced component lower band and a stochastic/noise component upper band based on MVF values. It also assumes a two-band mixed excitation that can handle the whole speech signal rather than only the excitation. Moreover, the variability of the harmonics features (amplitudes and phases) with respect to continuous F0 is very high; therefore, I decide to parameterize speech from the same parameters (F0, MVF, and MGC) used in the baseline. The novelty behind this vocoder is to use harmonic features to facilitate and improve the synthesizing step before speech reconstruction.

The performance of this system has been also evaluated through subjective listening test[8] (Figure 9). Experiments demonstrate that our proposed model generates higher output speech quality than the baseline (that is a source-filter based model). Moreover, it was found that the results obtained with the proposed vocoder were preferred over STRAIGHT vocoder. Table 4 also point out that the continuous vocoder has few parameters compared to the WORLD and STRAIGHT vocoders, and it is computationally feasible; therefore, it is suitable for real-time operation.

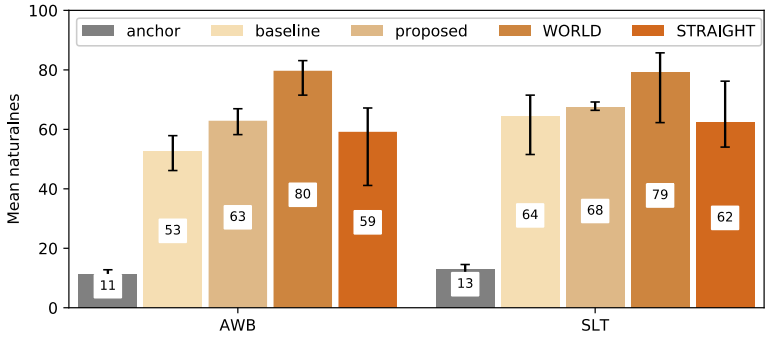


Figure 9. Results of the MUSHRA listening test for the naturalness question.

Table 4. Parameters 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
STRAIGHT	F0: 1 + Aperiodicity : 1024 + Spectrum: 1024	Mixed

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

Thesis IV.2: CSM with Deep Learning [C5] *Based on the above results, I built and developed a deep learning based bidirectional LSTM version of the continuous sinusoidal model to generate high-quality synthesized speech. I showed that the proposed framework converges faster and provides satisfactory results in terms of naturalness and intelligibility comparable to the high-quality WORLD model based TTS.*

Continuous vocoder based on sinusoidal model was designed to overcome shortcomings of discontinuity in the speech parameters and the computational complexity of modern vocoders. Moreover, the novelty behind this vocoder is to use harmonic features to facilitate and improve the synthesizing step before speech reconstruction, which was successfully outperform state-of-the-art vocoders performance (e.g. STRAIGHT) in synthesized speech. In this Thesis, I addressed and investigated the use of sequence-to-sequence modeling based TTS using the CSM to model contF0, MVF, and MGC for more natural sounding speech synthesis.

To ensure an optimal training process can be used to enhance the performance of the proposed vocoder based TTS, 4 feed-forward hidden layers each consisting of 1024 units followed by a single Bi-LSTM layer with 385 units, will be used in this work to train the CSM parameters. To demonstrate the efficiency of our proposed model, we performed a web-based MUSHRA listening test[9]. Consequently, it can be observed from Figure 10 that the proposed framework outperforms the baseline vocoder (Mann-Whitney-Wilcoxon ranksum test, with a 95% confidence level). It can be also seen that the difference between CSM and WORLD vocoders is not significant. This means that CSM based TTS is beneficial in the statistical deep recurrent neural networks and it almost reached the level of the state-of-the-art high quality vocoder.

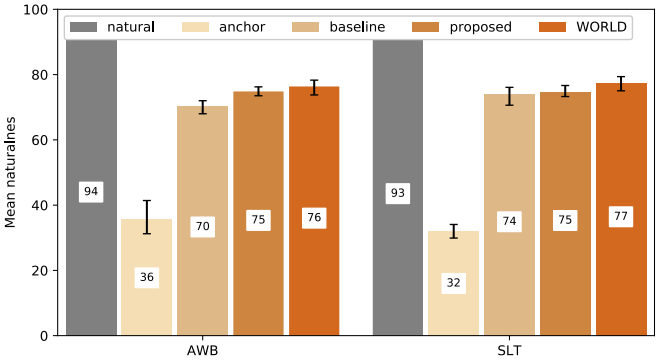


Figure 10. MUSHRA scores for the naturalness question.

[9] http://smartlab.tmit.bme.hu/ijcnn2019_vocoder

4.5. Thesis group V: Voice Conversion

Voice conversion (VC) aims to modify the speech signal of a source speaker into that of a target speaker. A well-designed VC system often consists of analysis, conversion, and synthesis modules. The process of parametrizing the input waveform into acoustic features and then synthesizing the converted waveform based on the converted features is one of the major factors that may degrade the performance of VCs. For this, the characteristics of the speech vocoder (analysis/synthesis system) given to the VC are of paramount importance.

I grouped the state-of-the-art vocoders based VC into three categories. a) Source-filter models: e.g., STRAIGHT [36]; b) Sinusoidal models: e.g., Harmonic plus Noise Model [37]; c) end-to-end complex models: e.g., adaptive WaveNet [38]. In the face of their clear differences, each model has advantages to work reasonably well, for a particular speaker or gender conversion task, which make them attractive to researchers. Even though these techniques achieve some improvements, the naturalness of the converted voice still deteriorates compared to the source speaker due to the over-smooth phenomenon or discontinuity problems, which makes the converted speech sound muffled. Therefore, improving the performance of converted voice is still a challenging research question. Nonetheless, such mismatch between the trained, converted, and tested features still exist, which often causes significant quality and similarity degradation. Consequently, simple and uniform vocoders, which would handle all speech sounds and voice qualities (e.g., creaky voice) in a unified way, are still missing in VC.

Moreover, traditional conversion systems focus on the prosodic feature represented by the discontinuous fundamental frequency assumption that depends on a binary voicing decision. Therefore, modelling of F0 in VC applications is problematic because of the differing nature of F0 observations between voiced and unvoiced speech regions. An alternative solution of increasing the accuracy of the acoustic VC model is using a continuous F0 to avoid alignment errors that may happen in voiced and unvoiced segments and can degrade the converted speech.

Therefore, the goal of the work reported in this Thesis was to propose a new VC model based on continuous parameters in both source-filter and sinusoidal models as shown in Figure 11. Listening test results show that the suggested models give state-of-the-art similarity results.

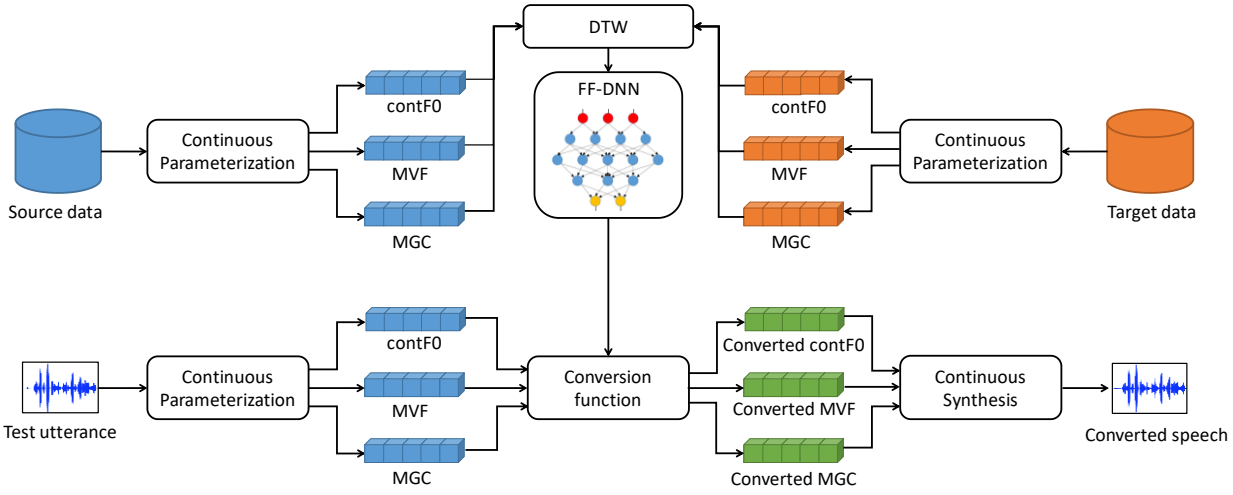


Figure 11. Flowchart of the proposed VC system.

Thesis V.1: Statistical VC with Source-Filter Model [C9, J3] *I proposed a novel VC system using the source-filter based continuous vocoder. I demonstrated that using continuous parameters provide accurate and efficient system that convert source speech signal to the target one. I experimentally proved that the new method improves similarity compared to the conventional method.*

Recently in SPSS, I have developed a novel continuous vocoder using contF0 in combination with MVF, which was shown to improve the performance under a FF-DNN compared to the HMM based TTS. The advantage of a continuous vocoder in this scenario is that vocoder parameters are simpler to model than in traditional vocoders with discontinuous F0. Here, I proposed computationally efficient and accurate model to achieve performance improvement consistently over the traditional techniques.

Unlike existing methods in the literature, the proposed structure implicates two major technical developments. First, I build a voice conversion framework that consists of a FF-DNN and a continuous vocoder to automatically estimate the mapping relationship between the parameters of the source and target speakers. Second, I apply a geometric approach to spectral subtraction to improve the signal-to-noise ratio of the converted speech and enhance anti-noise property of the proposed vocoder.

Listening tests showed very clear preference for the proposed method in mimicking the target speaker's speaking style and gives better performance as measured by objective evaluation and subjective listening[10] tests (Figure 12).

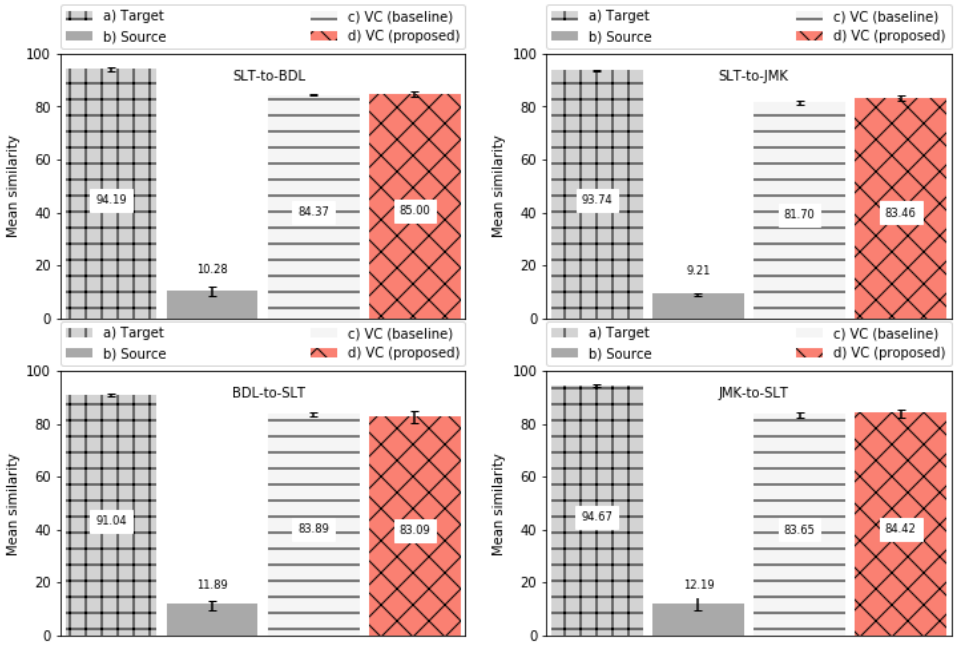


Figure. 12 MUSHRA scores for the similarity question. Errorbars show the bootstrapped 95% confidence intervals.

[10] http://smartlab.tmit.bme.hu/vc2019

Thesis V.2: Parallel VC with Sinusoidal Model [C7, J5] *I proposed a new approach to develop a voice conversion system using the continuous sinusoidal model, which decomposes the source voice into harmonic components and models contF0 to improve VC performance. I have validated the new system on parallel training data and showed its superiority against state-of-the-art solutions.*

The main challenge introduced in current voice conversion is the tradeoff between speaker similarity and computational complexity. Most of the existing VC techniques discard or does not typically preserve phase spectrum information which leads to the degradation of the performance of VC. However, the effectiveness of phase information in detecting synthetic speech has recently been proved by [39]. One possible way of enhancing the accuracy of VC models is to incorporate phase information to achieve superior synthesized speech. Therefore, I developed in this Thesis a sinusoidal type synthesis model based on contF0 for achieving high-quality converted speech.

Unlike conventional source-filter based techniques existing in the literature, the proposed structure use harmonic features to facilitate and improve the converted synthesizing step before speech reconstruction. Four speakers are chosen as the main corpus in this Thesis, and I conducted intra-gender and cross-gender pairs. Consequently, I ran 48 experiments in order to measure the performance of the proposed VC system. A more detailed case-by-case analysis by fwSNRseg and LLR are shown in Table 5. As a result, these findings demonstrate that the CSM can yield a good performance comparable to other systems (e.g., WORLD model). Moreover, the comparison of the continuous features of one speech frame converted by the proposed method are given in Figure 13. It may be observed that the converted parameters are more similar in general to the target one than the source one. Additionally, a perceptual listening[11] test was designed to test and evaluate the performance of the proposed model.

The MUSHRA similarity scores of the listening test are presented in Figure 14. An interesting note is that the listeners preferred our system compared to others developed earlier. This means that our proposed model has successfully converted the source voice to the target voice on the same-gender and cross-gender cases, which get higher scores in the MUSHRA test.

Table 5. Average scores on converted speech signal per each of the speaker pairs conversion. Higher value indicates better performance for fwSNRseg whereas lower value indicates better performance for LLR.

Model	WORLD		MagPhase		Sprocket		Proposed	
	fwSNRseg	LLR	fwSNRseg	LLR	fwSNRseg	LLR	fwSNRseg	LLR
BDL → JMK	2.19	1.57	3.21	1.37	2.20	1.48	2.47	1.50
BDL → SLT	1.12	1.72	1.25	1.69	1.04	1.49	2.33	1.57
BDL → CLB	0.79	1.83	1.65	1.72	0.37	1.69	1.66	1.74
JMK → BDL	1.31	1.76	2.49	1.56	1.73	1.63	2.15	1.57
JMK → SLT	0.55	1.74	1.93	1.56	0.11	1.64	1.54	1.65
JMK → CLB	1.45	1.74	1.75	1.66	0.69	1.60	1.81	1.67
SLT → BDL	1.65	1.71	1.60	1.70	1.80	1.51	2.95	1.49
SLT → JMK	2.16	1.61	2.71	1.42	0.713	1.56	2.59	1.39
SLT → CLB	1.51	1.75	2.89	1.59	2.32	1.56	2.51	1.50
CLB → BDL	0.97	1.81	1.60	1.70	0.95	1.72	1.92	1.60
CLB → JMK	2.50	1.49	2.74	1.40	0.98	1.46	3.00	1.30
CLB → SLT	0.98	1.70	2.17	1.53	1.96	1.54	2.12	1.47

[11] http://smartlab.tmit.bme.hu/sped2019_vc

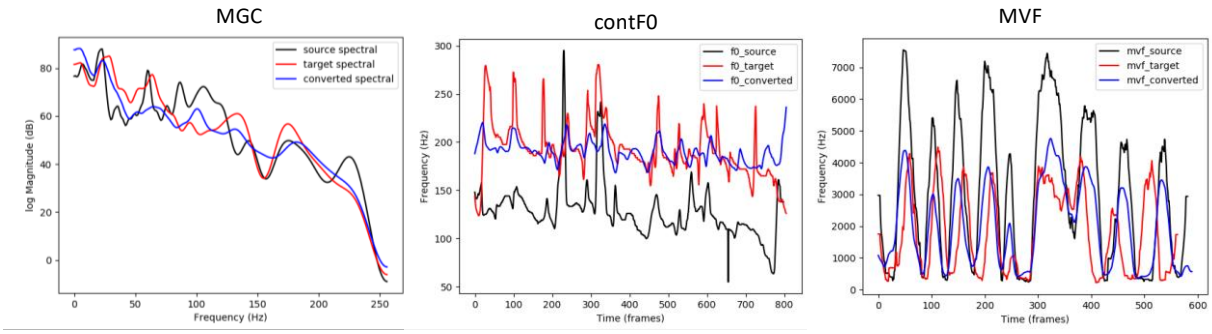


Figure 13. Example of the natural source (black), target (red), and converted (blue) spectral envelope, contF0, and MVF trajectories using the proposed method.

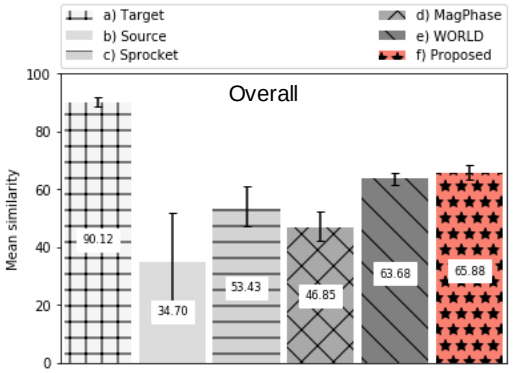


Figure 14. MUSHRA scores for the similarity question. Higher value means better overall quality.

5. Applicability of the results

This Thesis work and its results are not only scientifically evaluable; they are useful for the current state of the speech technology applications and profession. I have proposed and thoroughly examined the concept of vocoding for speech synthesis and voice conversion. Certainly, the proposed frameworks and algorithms are independent of any spoken language, which can directly be used in many speech applications to provide significantly better synthetic speech performance. Here, the practical applications of the results of this dissertation are summarized.

The results of Thesis group I – noise modelling – is expected to be used for making speech synthesis more natural and expressive. In particular, the temporal envelope of Thesis I.1 is a relevant acoustic feature which can contribute to get a reliable estimation of the speech features in human study, whereas it can be used to avoid artifacts near the voicing boundaries in order to improve the quality of statistical parametric speech synthesis system. The novel outcomes of Thesis I.2 play a role to improve speech intelligibility and enhance voice qualities (hoarseness, breathiness, and creaky voice) in various speech synthesis systems; for example, creaky voice segments are not properly reconstructed in both HMPD and STRAIGHT vocoders [14]. Thus, Theses group I attempted to further assist in reducing the effects of residual noise caused by the inaccurate excitation modeling.

The results of Thesis group II – excitation Modelling – provides a reference for selecting appropriate techniques to optimize and improve the performance of current fundamental frequency estimation methods based on speech synthesis. Thesis II.1 can be used to reduce the fine error that the voiced section is wrongly identified as the unvoiced section, and improving temporal resolution of the estimated F0 trajectory. The time warping scheme has the ability to track the time-varying F0 period, and reduce the amount of F0 trajectory deviation from their harmonic locations. Whereas the instantaneous frequency approach is computationally inexpensive and can be highly useful in real-time processing speech synthesis applications. Thesis II.2 can be used to indicate the degree of voicing in the excitation, to detect the pitch exactly in various speech applications, and hence subsequently reducing the influence of buzziness caused by the vocoder.

The results of Thesis group III – acoustic modelling – is the possibility of creating new DNN-TTS voices automatically (e.g., from a speech corpus of telephone conversations) for simple devices (e.g. smartphones). This Thesis demonstrates the superiority of DNN acoustic models over the decision tree used in the HMM. Thesis II.1 based on DNN and Thesis II.2 based on RNN have already been applied in TTS with a developed vocoder as a simple, flexible, and powerful alternative acoustic model for SPSS to significantly improve the synthetic speech quality.

The results of Thesis group IV – sinusoidal modelling – introduces a vocoder-based speech synthesis system to improve the sound quality of real-time applications. This new speech synthesis system can be used in various speech technology, such as voice conversion, speech manipulation, and singing synthesizers. Thesis IV.1 can handle a wide variety of speakers and speaking conditions and give natural sounding speech comparable to the state-of-the-art STRAIGHT vocoder. Besides significant quality improvements over the baseline, the resulting system in Thesis IV.2 can be used in many speech applications, including message readers (SMS, e-mail, e-book, etc.), and navigation systems.

The results of Thesis group V – voice conversion – give an advanced novel approach to improve the conversion performance. The systems from both Thesis V.1 and V.2 have already been applied in the speaker conversion application using the continuous vocoder based on source-filter and sinusoidal model, respectively. These methods were tested with English speech corpora; however, it could be easily extended to other languages as well. This Thesis can also be applied in emotion conversion, virtual-augmented reality systems (voice avatars), accent conversion in language learning [40], and other speech assistance for overcoming speech impairments [41].

In addition to the individuals mentioned above, such an application is already under development within cooperation with an Egyptian university to create Arabic text-to-speech synthesis engine, in which continuous vocoder was applied on a modern standard Arabic audio-visual corpus which is annotated both phonetically and visually to produce a high-quality Arabic emotion TTS system [J1]. The general application of this TTS engine is to make a screen reader for Arabic's blind users. Moreover, continuous vocoder has already been applied in silent speech interfaces [C6], which is a form of spoken communication where an acoustic signal is not produced [42]. Continuous parameters were predicted form ultrasound tongue image by using the automatic articulatory-to-acoustic mapping, in which deep convolution neural network was used to learn the mapping task. Such an application can be applied to help the communication of the speaking impaired (e.g. patients after laryngectomy).

6. References

[1] Heiga Zen, Keiichi Tokuda, Alan W. Black, "Statistical parametric speech synthesis," *Speech Communication,* vol. 51, no. 11, pp. 1039-1064, 2009.

[2] Kawahara H., Masuda-Katsuse I., de Cheveigne 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, pp. 187-207, 1999.

[3] Mohammadi SH, Kain A, "An overview of voice conversion systems," *Speech Communication,* vol. 88, pp. 65-82, 2017.

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

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

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

[7] Tokuda K., Kobayashi T., Masuko T., Imai S., "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," *Proc. ICSLP,* pp. 1043-1046, 1994.

[8] Imai S., Sumita K., 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.

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

[10] Quackenbush S., Barnwell T., Clements M., Objective Measures of Speech Quality, Englewood Cliffs: NJ: Prentice-Hall, 1988.

[11] Ma J., Hu Y., 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.

[12] Kubichek R.F., "Mel-cepstral distance measure for objective speech quality assessment," *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing,* vol. 1, pp. 125-128, 1993.

[13] Rabiner L. R., Cheng M. J., Rosenberg A. E., McGonegal C. A., "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 24, no. 5, pp. 399-417, 1976.

[14] Degottex G., Erro D., "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP, Journal on Audio, Speech, and Music Processing,* vol. 38, no. 1, pp. 1-16, 2014.

[15] Morise M., Watanabe Y., "Sound quality comparison among high-quality vocoders by using re-synthesized speech," *Acoustical Science and Technology,* vol. 39, no. 3, pp. 263-265, 2018.

[16] Espic F., Valentini-Botinhao C., King S., "Direct Modelling of Magnitude and Phase Spectra for Statistical Parametric Speech Synthesis," *in INTERSPEECH 2017,* pp. 1383-1387, 2017.

[17] Degottex G., Lanchantin P., Gales M., "A Log Domain Pulse Model for Parametric Speech Synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 1, pp. 57-70, 2018.

[18] Kobayashi K., Toda t., "sprocket: Open-Source Voice Conversion Software," *Proceedings of the Odyssey: The Speaker and Language Recognition,* pp. 203-210, 2018.

[19] Kawahara H., Agiomyrgiannakis Y., and Zen H., "Using instantaneous frequency and aperiodicity detection to estimate f0 for high-quality speech synthesis," in *9th ISCA Workshop on Speech Synthesis*, CA, USA, 2016.

[20] Hua K., "Improving YANGsaf F0 Estimator with Adaptive Kalman Filter," in *Proc. of the Interspeech*, Stockholm, Sweden, 2017.

[21] Zhizheng W., Watts O., King S., "Merlin: An Open Source Neural Network Speech Synthesis System," in *Proceeding 9th ISCA Speech Synthesis Workshop (SSW9)*, Sunnyvale, CA, USA, 2016.

[22] Boersma P., "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341-345, 15 November 2002.

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

[24] Hu Q., Richmond K., Yamagishi J., Latorre J., "An experimental comparison of multiple vocoder types," in *Proc. ISCA SSW8*, pp.155–160, 2013.

[25] Erro D., Sainz I., Navas E., 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.

[26] Najafabadi M., Villanustre F., Khoshgoftaar T., Seliya N. , Wald R., Muharemagic E., "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2:1, pp. 1-21, 2015.

[27] Csapó T.G., Németh G., Cernak M., "Residual-Based Excitation with Continuous F0 Modeling in HMM-Based Speech Synthesis," *3rd International Conference on Statistical Language and Speech Processing*, vol. 9449, pp. 27-38, 2015.

[28] Zen H., Senior A., Schuster A., "Statistical parametric speech synthesis using deep neural network," *Proc. ICASSP*, pp. 7962-7966, 2013.

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

[30] Li Q., Mark R.G., Clifford G.D., "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter," *Physiological Measurement*, vol. 29, no. 1, pp. 15-32, 2008.

[31] Kumaresan R., Ramalingam C.S., "On separating voiced-speech into its components," *Proc. of the 27th Asilomar Conferance Signals, Systems, and Computers*, Vols. 1-2, pp. 1041-1046, Pacific Grove, CA, 1993.

[32] Flanagan J.L., Golden R.M., "Phase vocoder," *The Bell System Technical Journal*, vol. 45, no. 9, pp. 1493-1509, 2009.

[33] McAulay R.J., Quatieri T.F., "Sinusoidal coding," *Speech Coding and Synthesis, Elsevier*, pp. 121-173, 1995.

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

[35] Hu, Q., Stylianou, Y., Maia, R., Richmond, K., and Yamagishi, J., "Methods for applying dynamic sinusoidal models to statistical parametric speech synthesis," *IEEE ICASSP*, pp. 4889-4893, South Brisbane, 2015.

[36] Toda T., Saruwatari H., Shikano K., "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," *Proceedings of the ICASSP*, pp. 841-844, 2001.

[37] Lifang W., Linghua Z., "A Voice Conversion System Based on the Harmonic Plus Noise Excitation and Gaussian Mixture Model," *Proceedings of the Instrumentation, Measurement, Computer, Communication and Control*, pp. 1575-1578, 2012.

[38] Sisman B., Zhang M., Sakti S., Li H., Nakamura, S., "Adaptive WaveNet Vocoder for Residual Compensation in GAN-Based Voice Conversion," *IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[39] Saratxagaa I., Sanchez J., Wu Z., Hernaeza I., "Synthetic Speech Detection Using Phase Information," *Speech Communication*, vol. 81, pp. 30-41, 2016.

[40] Daniel Felps, Heather Bortfeld, Ricardo Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech communication*, vol. 51, no. 10, pp. 920-932, 2009.

[41] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134-146, 2012.

[42] Denby B., Schultz T., Honda K., Hueber T., Gilbert J.M., Brumberg J.S., "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270-287, 2010.

7. Publication

7.1. Publications related to Ph.D. Thesis

International journals (peer-reviewed)

[J1] Mohammed Salah Al-Radhi, Omnia Abdo, Tamás Gábor Csapó, Sherif Abdou, Géza Németh, Mervat Fashal, A continuous vocoder for statistical parametric speech synthesis and its evaluation using an audio-visual phonetically annotated Arabic corpus, *Computer Speech and Language*, ScienceDirect Elsevier, 60, pp. 1-15, 2020. (WoS, IF = 1.86, Q1), [50% · 6p = 3 points]

[J2] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, Adaptive refinements of pitch tracking and HNR estimation within a vocoder for statistical parametric speech synthesis. *Applied Sciences*, 9, 2460, pp. 1-23, 2019. (WoS, IF = 2.22, Q1), [50% · 6p = 3 points]

[J3] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, Continuous vocoder applied in deep neural network based voice conversion, *Multimedia Tools and Applications*, 78 (23), Springer, pp. 1-24, 2019. (WoS, IF = 2.10, Q1), [50% · 6p = 3 points]

[J4] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, Continuous noise masking based vocoder for statistical parametric speech synthesis, *IEICE Transactions on Information and Systems*, accepted, E103-D, 05, 2020. (WoS, IF = 0.58, Q3), [50% · 6p = 3 points]

[J5] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, Noise and acoustic modeling in text-to-speech synthesis and parallel voice conversion, *Romanian Journal of Information Science and Technology*, submitted on 26th Feburary 2019, "Under Review". (WoS, IF = 0.66, Q3), [50% · 6p = 0 points]

International conferences (peer-reviewed)

[C1] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, Time-domain envelope modulating the noise component of excitation in a continuous residual-based vocoder for statistical parametric speech synthesis, *in Proceedings of the 18th Interspeech conference*, pp. 434-438, Stockholm, Sweeden, 2017. (Scopus), [50% · 3p = 1.5 points]

[C2] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, Continuous vocoder in feed-forward deep neural network based speech synthesis, *in Proceedings of the 11th Digital speech and image processing conference*, pp. 1-4, Novi Sad, Serbia, 2017. (SemanticScholar), [50% · 3p = 1.5 points]

[C3] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, Deep recurrent neural networks in speech synthesis using a continuous vocoder. In: Karpov A., Potapova R., Mporas I. (eds) *Speech and Computer. SPECOM. Lecture Notes in Computer Science*, vol 10458. Springer, pp. 282-291, Hatfield, England, 2017. (Scopus, chapter book), [50% · 3p = 1.5 points]

[C4] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, A continuous vocoder using sinusoidal model for statistical parametric speech synthesis. In: Karpov A., Jokisch O., Potapova R. (eds) *Speech and Computer. SPECOM. Lecture Notes in Computer Science*, vol 11096. Springer, pp. 11-20, Leipzig, Germany, 2018. (Scopus, chapter book), [50% · 3p = 1.5 points]

[C5] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, RNN-based speech synthesis using a continuous sinusoidal model, *in Procedings of the 28th IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, Budapest, Hungary, 2019. (IEEE), [50% · 3p = 1.5 points]

[C6] Tamás Gábor Csapó, Mohammed Salah Al-Radhi, Géza Németh, Gábor Gosztolya, Tamás Grósz, László Tóth, Alexandra Markó, Ultrasound-based silent speech interface built on a continuous vocoder, *in Proceedings of the 20th Interspeech conference*, pp. 894-898, Graz, Austria, 2019. (Scopus), [17% · 3p = 0.51 points]

[C7] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh, Parallel voice conversion based on a continuous sinusoidal model, *in Proceedings of the 10th IEEE Speech Technology and Human-Computer Dialogue conference*, pp. 1-6, Timisoara, Romania, 2019. (IEEE), [50% · 3p = 1.5 points]

International abstract conferences (peer-reviewed)

[C8] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, Effects of adding a Harmonic-to-Noise Ratio parameter to a continuous vocoder, *in Proceedings of the 6th of the UKspeech*, Cambridge University, England, 2017. (poster, 1 page), [50% · 1p = 0.5 points]

[C9] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, High quality continuous vocoder in deep recurrent neural network based speech synthesis, *in Eastern European Machine Learning*, Bucharest, Romania, 2019. (poster, 2 pages), [50% · 0p = 0 points]

[C10] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh, Improving continuous F0 estimator with adaptive time-warping for high-quality speech synthesis, *in Beszédkutatás (conference of the speech reseacrch)*, Budapest, Hungary, 2018. (oral, 2 pages), [50% · 0p = 0 points]

International doctoral consortium conferences (peer-reviewed)

[C11] Mohammed Salah Al-Radhi, High quality continuous residual-based vocoder for statistical parametric speech synthesis, *International Speech Communication Association (ISCA-SAC), Interspeech*, KTH Royal Institute of Technology, Stockholm, Sweeden, 2017.
(Googlescholar, oral, 3 pages), [100% · 1p = 1 points]

[C12] Mohammed Salah Al-Radhi, High-quality vocoding for speech synthesis and voice conversion, *International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2019.
(oral, 3 pages), [100% · 0p = 0 points]

7.2. Additional publications (my contribution related to deep learning)

International journals (peer-reviewed)

[J6] Waleed I. Hameed, Baha A. Sawadi, Safa J. Al-Kamil, Mohammed Salah Al-Radhi, Yasir I. Al-Yasir, Ameer L. Saleh, Raed A. Abd-Alhameed, Prediction of solar irradiance based on artificial neural networks. *Inventions*, 4, 45, pp. 1-10, 2019. (Scopus), [15% · 6p = 0.9 points]

7.3. Independent citations

[C4-1] Jiang C., Chen Y., Cho C., A Novel Genetic Algorithm for Parameter Estimation of Sinusoidal Signals, *12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Suzhou, pp. 1-5, 2019.

Total publication score: 23.91 points