A Cross-vocoder Study of Speaker Independent Synthetic Speech Detection using Phase Information

Jon Sanchez[1], Ibon Saratxaga[1], Inma Hernaez[1], Eva Navas[1], Daniel Erro[1,2]

[1] Aholab Signal Processing Laboratory, University of the Basque Country UPV/EHU, Spain
[2] Ikerbasque, Basque Foundation for Science, Bilbao, Spain.

ion@aholab.ehu.es, ibon@aholab.ehu.es, inma@aholab.ehu.es, eva@aholab.ehu.es, derro@aholab.ehu.es

Abstract

Current speaker verification systems are vulnerable to advanced speech manipulation techniques such as voice conversion and speaker adaptation for TTS systems. Effective anti-spoofing systems that allow the discrimination between human and synthetic impostors have been developed. However, many of them still present two main drawbacks: speaker dependency and, more importantly, counterfeiting technique dependency. Thus, getting a universal synthetic speech detector (SSD) remains an open issue. This paper explores the feasibility of such a system using a statistical classifier for human and synthetic speech. Provided the great diversity of counterfeiting techniques, we have chosen to model a variety of state-of-the-art minimum-phase vocoders, creating imposter synthetic signals by copy-synthesis. Two speech parameter sets are used: MFCCs as a canonical baseline and relative phase shift (RPS) based parameterization. Phase related parameters allow synthetic speech detection based on the presumably different phase structures of the human and synthetic signals due to the fact that most speech synthesis and conversion techniques disregard phase information. The results of the experiments show that speaker independent classifiers perform very well for every vocoder. Cross-vocoder experiments show that the system is highly dependent on the type of vocoder, and that RPS parameterization performs better than MFCC for multi-vocoder models.

Index Terms: synthetic speech detection, phase information, anti-spoofing

1. Introduction

In the last few years, voice has become one of the most important biometric parameter for people identification purposes: it can be used to check if a claimed identity actually corresponds with the related voice. The task is called Speaker Verification (SV) [1].

Nowadays speech synthesis and voice conversion have experienced a great improvement [2][3][4], and, since it is no longer necessary to gather big amounts of data to get a quality adapted synthetic voice, it could be easy to fake a person's voice in order to gain access to a speaker verification controlled system. To be able to detect these faked voices two main strategies have been proposed. The first one implies improving the verification system itself, usually changing the modeling technique or the parameters used, so that both human and synthetic impostors can be detected and blocked [5][6][7]. The second strategy consists on building a separate synthetic speech detection (SSD) module to be used before or after the SV system. The SSD module implements specific parameters and detection techniques focused on the presumed differences of synthetic speech: pitch variations [8], interframe statistical similarities in some parameters [9][10], phase information [11][12] , temporal modulation [13], etc.

Some of these SSD systems apply a speaker dependent approach comparing the discriminating features against the expected ones for a specific speaker [10][11]. Their main disadvantage is that it is necessary to develop a specific synthetic model for each speaker allowed into the system. A second approach seeks for speaker independent SSD modules which attempt to detect synthetic voices from unknown speakers not present in the training signals [8][9][12][13]. This approach is very convenient as it avoids creating new SSD models every time a new speaker is added to the system acceptance list.

Most of the published SSD systems address the modeling of the synthetic signals by considering a single counterfeiting technique (i.e. voice conversion, synthetic speaker adapted voice, etc.). A more general approach consists on using copy-synthesis to create imposter samples, aiming for a model capable of classifying any signal generated with that synthesis method, whichever the specific conversion or adaptation technique might be. This approach avoids the need to obtain the speaker adapted TTS systems or the target speaker conversion functions. In [11] it is concluded that copy-synthesis detection poses a more difficult task for the SSD compared to the detection of a specific attacking technique, and thus, it is a good approach to test the limits of the classifiers. This is also the approach followed in this work.

In any case, to our knowledge, no studies have been published about the performance of these systems when there is no previous knowledge of the synthesis method used on the spoofing attack. Many synthesis and voice conversion methods are based on vocoders, so vocoder dependency is a real problem in practical situations, where prior knowledge of the spoofing attack technique cannot be assumed.

In the present work we study the performance of a SSD system with speaker independent modeling for different state-of-the-art vocoders which are widely used in statistical speech synthesis and voice conversion. We also use two different parameter sets to generate the models: canonical module based MFCC [14] and phase based Relative Phase Shifts (RPS) [15]. As explained in [11], most popular vocoders do not make use of the phase information, so the phase differences between an original signal and a resynthesized one can be relevant. The RPS representation has been successfully used in [11] and [16] in the context of speaker dependent models.

Furthermore, we analyze the feasibility of getting a vocoder independent SSD system, studying the cross-vocoder performance and the problems of multi-vocoder modeling.

The paper is organized as follows. First the whole system is described. Then, the evaluation experiment is outlined and

the results are presented. Finally, some conclusions are drawn.

2. System Description

2.1. General Architecture

Figure 1 shows the general architecture of the SSD system. Its purpose is to take a decision about the synthetic nature of the input speech signal. If previous knowledge of the speaker identity is not necessary, the SSD module can be inserted before or after the traditional SV system.

During the training phase, 512 mixtures GMM models for both natural speech (λ_{human}) and synthetic speech (λ_{synth}) were created, using two different sets of parameter vectors: phase-based vectors and MFCC vectors. The synthetic signal will be a counterfeit of the original speaker voice. As mentioned in the previous section, in this work we use coded-decoded (or copy-synthesized signals) generated using several state-of-the art vocoders, as convenient substitutes of the possible TTS or converted signals.

To perform the synthetic speech detection task, the system will test a candidate vector sequence **Y** of length N against both natural speech and synthetic speech models to get the corresponding likelihood values $p(\mathbf{Y}|\lambda_{human})$ and $p(\mathbf{Y}|\lambda_{synth})$. Then according to (1) the log likelihood ratio Λ is calculated, taking the candidate as human if it exceeds a certain decision threshold θ which was set to the Equal Error Rate (EER) point in the experiments.

$$\Lambda(\mathbf{Y}) = \log p(\mathbf{Y}|\lambda_{human}) - \log p(\mathbf{Y}|\lambda_{synth}) \quad (1)$$

where

$$\log p(\mathbf{Y}|\lambda) = \frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{y}_n|\lambda) \quad (2)$$

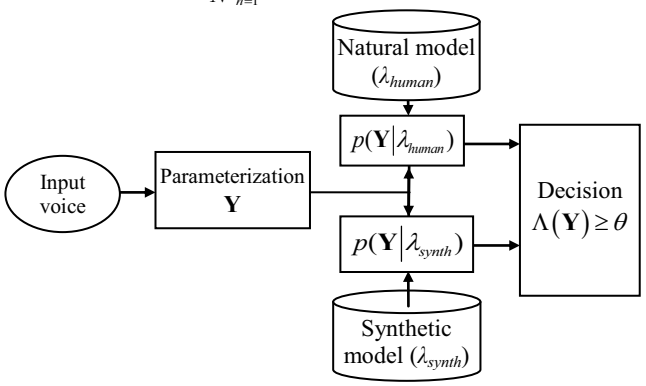

Figure 1: *SSD system structure.*

2.2. Parameterization

For every signal used two different sets of parameters have been obtained: the RPS parameters, and the MFCC parameters that will be used as a baseline.

2.2.1. RPS parameters

The RPS is a representation for the harmonic phase information described in [15]. Harmonic analysis models each frame of a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency.

$$h(t) = \sum_{k=1}^{N} A_k \cos\left(\varphi_k(t)\right) \qquad \varphi_k(t) = 2\pi k f_o t + \theta_k \quad (3)$$

where N is the number of bands, A_k the amplitudes, $\varphi_k(t)$ the instantaneous phase, f_0 the pitch or fundamental frequency and θ_k the initial phase shift of the k-th sinusoid. The RPS representation consists in calculating the phase shift between every harmonic and the fundamental component (k=1) at a specific point of the fundamental period, namely the point where $\varphi_o=0$.

$$\psi_k(t_a) = \varphi_k(t_o) = \varphi_k(t_a) - k\varphi_1(t_a) \quad (4)$$

Equation (4) defines the RPS transformation which allows computing the RPSs (ψ_k) from the instantaneous phases at any point (t_a) of the signal. The RPS values are wrapped to the [-π, π] interval.

The RPS values are not suitable for statistical modelling, so to create and test the models the so-called DCT-mel-RPS parameterization is used instead. These parameters, thoroughly explained in [17], have produced good results in other tasks where statistical modelling is used, such as ASR, Speaker Identification and also Synthetic Speech Detection tasks. To obtain the parameters, the differences of the unwrapped RPS values are filtered with a mel filter bank (48 filters) and a discrete cosine transform (DCT) is applied to the resulting sequence. The DCT is truncated to 20 values and the Δ and ΔΔ values are calculated. The averaged value of the slope of the unwrapped RPS values is also included which leads to a total of 63 phase-based parameters, calculated only for voiced frames, every 10ms.

2.2.2. MFCC parameters

To be used as reference or baseline, MFCC parameters [14] are also calculated and modeled. 13 MFCC parameters are used, not including MFCC-0. Their Δ and ΔΔ values are also computed, which leads to 39 parameters calculated every 10 ms in the segments of the signal where there is voice activity.

2.3. Vocoders

Three different vocoders have been used:

- The basic **MLSA** vocoder included in the HTS demo release [18][19]. During the analysis it estimates the fundamental frequency and performs Mel-cepstral analysis [20] (order 24 for f_s=16kHz). The reconstructed waveform is built by filtering a simple F0-dependent pulse/noise excitation through the so called MLSA filter [14], which is related to the Mel-cepstral coefficients. This basic MLSA vocoder has been used for vocoded speech detection in [12].
- The **STRAIGHT**-based vocoder included in the HTS demo release [18][21]. STRAIGHT is a high-quality speech analysis, manipulation and reconstruction tool that represents the speech signal by means of its fundamental frequency, a high-resolution spectral envelope, and an aperiodicity frequency contour [22]. In statistical parametric systems [21], the spectral envelope is typically translated into a Mel-cepstral representation (order 39 for f_s=16kHz), while the aperiodicity values are averaged within 5 specific bands. The STRAIGHT vocoder was used for synthetic speech detection in [11][13]. It is widely used for voice conversion and speech synthesis.
- **AHOCODER,** a recently proposed vocoder based on the harmonics plus noise model (HNM) [23][24]. It parameterizes speech into three different streams (namely fundamental frequency, Mel-cepstral coefficients (order

39 for f_s=16kHz) and maximum voiced frequency) and uses HNM-related procedures for signal analysis and reconstruction. It has been applied to both speech synthesis [23][24] and voice conversion [3].

Experiments were made by encoding the recorded signals by means of these three vocoders. Previous to the coding step, the DC component of all the signals was filtered out, and their energy normalized. This was done because not all the coders treat the DC component and the energy in the same way. Also, the polarity of the signals was homogenized, as required by the RPS parameterization [25].

All these vocoders use minimum phase approaches during waveform reconstruction. We have to mention that some recently proposed vocoders do not make use of this technique [26][27][28][29], and thus, the robustness of the phase-based SSD system against them should be studied in future works.

2.4. Database

The WSJ database [30] has been used, as in [9] and [13]. Its main advantages are the large number of speakers (283) and the high S/N ratio of the recordings, which is important to get high quality synthetic signals.

From the SI-284 set, 8599 sentences were randomly selected, containing speech from all the 283 speakers. All the sentences were copy-synthesized using the three vocoders, thus getting a total of 8599×4 signals. Every signal, original and synthetic, has been downsampled to 8kHz, and then parameterized. To evaluate the performance of the system when a different amount of speakers is used to build the models, three different sets were defined:

- 30-SPK: 30 randomly selected speakers (about 900 sentences) are used to build the models. The remaining 253 speakers form the test set (about 7700 sentences).
- 50-SPK 50 randomly selected speakers (about 1500 sentences) are used to build the models. The remaining 233 speakers form the test set (about 7100 sentences).
- 140-SPK 140 randomly selected speakers (about 4200 sentences) are used to build the models. The remaining 143 speakers form the test set (about 4400 sentences).

3. Experiments and Results

3.1. Speaker independency

In order to evaluate the performance of the speaker independent system, signals from different sets of speakers have been used to create the human and synthetic speech models. The three sets of training data defined in Section 2.4 have been used to study the influence of the number of speakers in the performance of the system.

The classifier has been evaluated separately using synthetic signals created with the aforementioned vocoders. Moreover, each test has been repeated choosing different speakers to form the train and test sets, in order to cross-validate the results. The figures presented in Table 1 and Table 2 are mean values of the EERs of the different repetitions. The threshold has been established independently at the EER for every classifier. In all tables, A stands for AHOCODER test signals, S for STRAIGHT, and M for MLSA.

Tables 1 and 2 show that both parameter sets get low error rates in the classification task. Although direct comparison is not possible between experiments, the results rank among the lowest EERs for such SSD published in the literature. RPS based SSD consistently provides good results (well below 1%) for any vocoder. MFCC based classifier gets very good results for some vocoders (e.g. MLSA and STRAIGHT) although it seems to fail with AHOCODER.

Table 1. *Average EER and standard deviation with RPS parameterization.*

	30-SPK	50-SPK	140-SPK
A	0.60 (σ=0.16)	0.51 (σ=0.20)	0.46 (σ=0.14)
S	0.62 (σ=0.19)	0.42 (σ=0.27)	0.15 (σ=0.09)
M	0.76 (σ=0.31)	0.79 (σ=0.31)	0.61 (σ=0.53)

Table 2. *Average EER and standard deviation with MFCC parameterization.*

	30-SPK	50-SPK	140-SPK
A	5.93 (σ=0.75)	5.60 (σ=0.57)	4.86 (σ=0.67)
S	0.25 (σ=0.13)	0.13 (σ=0.08)	0.10 (σ=0.07)
M	0.02 (σ=0.01)	0.01 (σ=0.01)	0.08 (σ=0.13)

In most of the experiments the evolution of the results shows slight improvements as the number of speakers used for training increases. Being these variations small, we can assume that speaker independent SSD can be built with any of the sets tested. In the following experiments we will use the SSD with models of 50 speakers.

3.2. Vocoder dependency

The first experiment to analyze the vocoder dependency of the SSD system consists in the cross-vocoder detection rate evaluation. Using the 50 speaker training set, the classifier trained with synthetic models of each vocoder (AHOCODER, STRAIGHT and MLSA) was tested with signals created with every other vocoder. The aim of the experiment is to find if a model created with one given vocoder is able to detect a synthetic signal generated by a different one. We have added a testing set (denoted by T in the tables) that includes the synthetic signals generated with all three vocoders and the same number of natural signals. This would emulate the case where the actual spoofing vocoder is not known beforehand.

Again, both RPS and MFCC parameterizations are used, and each combination repeated five times using different 50-SPK sets to cross-validate the results. The figures presented in Table 3 are mean values of the EERs.

Table 3. *Average EER value for the cross vocoder SSD with RPS and MFCC parameters.*

	AHOCODER model		STRAIGHT model		MLSA model	
	RPS	MFCC	RPS	MFCC	RPS	MFCC
A	0.51	5.60	2.80	25.59	11.17	50.73
S	2.60	20.27	0.42	0.13	1.54	13.71
M	27.04	83.55	26.66	0.93	0.79	0.01
T	12.53	36.02	11.80	14.00	7.07	27.46

The shaded values in Table 3 show the EER for matching synthetic models and test signals, previously shown in Tables 1 and 2 for 50-SPK models. The performance of the cross-vocoder detection falls dramatically compared to the matched vocoder results, with few exceptions where the difference is smaller: for the RPS parameterization AHOCODER and STRAIGHT get better results when being cross-tested that

those from MLSA. In any case, it can be said that the system is clearly vocoder dependent.

Comparing the RPS results with the baseline MFCC system, the results from Table 3 show that the RPS parameterization works better in most cases for the task of detecting synthetic speech using models from a different vocoder, suggesting that the MFCC parameterization is more vocoder specific than the RPS one. Consequently, the RPS results are always better when we take into account tests involving all kinds of vocoded signals (T test set).

3.3. Multi-vocoder model evaluation

As seen in the previous experiment, a model from a specific vocoder is generally unable to correctly classify a synthetic signal created with another one. Thus, it is worth exploring if it is possible that a model created using signals from different vocoders can generalize and detect synthetic speech created with a foreign system.

With that aim, models with combinations of two different vocoders have been created, and then tested with the three kinds of synthetic signals against the human ones. Additionally, a model with the three vocoders has been trained and tested with every kind of signal.

Again, both RPS and MFCC parameterizations are applied using 50 speaker models, and each combination repeated five times using different speakers to cross-validate the results.

Results in Table 4 show that the multi-vocoder model cannot improve the detection level of the vocoder specific model, when the vocoder used in the spoofing attack is known. For instance, when using the STRAIGHT model to detect STRAIGHT synthesized counterfeits, the EER is 0.42% and 0.13% for RPS and MFCC respectively. If MLSA signals are added to the model, the EER raises to 0.70% for the RPS parameterization and to 2.22% for MFCC. As the example shows, this drop of the results is different for the two parameterization methods: while RPS can integrate new vocoder information in a single model with an approximate error rise of about 70% for the different vocoders, for MFCC parameters the error increase is higher.

Table 4. *Average EER for the multi vocoder SSD with RPS and MFCC parameters.*

	A & S model		S & M model		A & M model	
	RPS	MFCC	RPS	MFCC	RPS	MFCC
A	0.92	8.67	4.70	41.34	1.12	27.61
S	0.75	0.42	0.70	2.22	1.47	8.81
M	25.32	3.09	0.95	0.03	1.19	0.02
T	11.39	5.42	2.95	22.92	1.32	17.33

Regarding the results for the non-matched vocoders, generally speaking RPS parameterization can slightly benefit from the introduction of new vocoders into the models. In the case of MFCC parameters this does not happen. Instead, the aggregation of vocoders leads to detection rates in between the detection rates of the two original vocoders. For instance, when using the STRAIGHT model to detect MLSA vocoded signals, the EER is 26.66% and 0.931% using RPS and MFCC respectively. If AHOCODER information is added to the model, the EER lowers to 25.32% for RPS but increases to 3.09% for the MFCC.

The resulting outcome of aggregating vocoders in a multi-vocoder system can be compared with the mono-vocoder option in the all-vocoder test set (T) (Tables 3 and 4). For the RPS parameterization the aggregation of vocoders performs better than the single models, that is, the worsening of the matched vocoder results is overcome by the improvement of the cross-vocoder results. This is not always the case for MFCC parameters, which, in addition, perform worse than RPS in most of the vocoder combinations for this test set.

Table 5. *Average EER value with models of the three vocoders, with RPS and MFCC parameters.*

	3 vocoders	
	RPS	MFCC
A	1.16	27.05
S	0.99	2.02
M	1.32	0.03
T	1.22	16.03

Table 5 shows the results achieved when using information of the three vocoders to create the models. Again, the RPS parameterization obtains uniform low error rates, for every kind of input signal with the exception of MLSA where MFCC gets a very low error rate.

4. Conclusions and Future Work

The experiments allow us to conclude that vocoded synthetic speech can be successfully detected using speaker independent models specifically created for that vocoder. The models can be trained using a moderate amount of speakers, and the design leads to very good results using RPS parameterization, while the performance of MFCC parameterization varies depending on the vocoder.

The classifiers built in this work are highly dependent on the vocoder: the good results obtained for the matched vocoder tests fall when trying to detect synthetic voice created with a given vocoder, applying a model trained with a different one. This happens using both RPS and MFCC parameterizations, but RPS seems to be more capable of generalizing, getting better results.

Aggregation of several vocoders into the models leads to small improvements especially when using RPS. Again, phase information appears to have greater extrapolation abilities than MFCC parameters. This is particularly interesting when there is no prior knowledge of the attacking vocoder: if the vocoder of the input signal is known, a vocoder specific model is best, but if it is unknown, an integrated model works better than the wrong model, when using RPS parameterization.

With these results, using a model with information of every available vocoder and RPS parameterization could achieve some success when detecting a spoofing attack performed with a vocoder unknown to the system. Nevertheless, a practical vocoder independent universal SSD, will require a more extensive future work.

5. Acknowledgements

This work has been partially supported by the Basque Government (Ber2Tek Project, IE12-333) and the Spanish Ministry of Economy and Competitiveness (SpeechTech4All project, TEC2012-38939-C03-03).

6. References

[1] Campbell, J. P., "Speaker recognition: A tutorial", Proc. IEEE, Cilt 85, pp. 1436-1462, 1997.

[2] Stylianou, Y., Cappe, O. and Moulines, E., "Continuous Probabilistic Transform for Voice Conversion", IEEE Transactions on Speech and Audio Processing, Vol. 6, No. 2, pp. 131-142, 1998.

[3] Erro, D., Navas, E. and Hernaez, I., "Parametric Voice Conversion based on Bilinear Frequency Warping plus Amplitude Scaling", IEEE Transactions on Audio, Speech, and Language Processing, vol. 21(3), pp. 556-566, 2013.

[4] Tokuda, K., Nankaru, Y., Toda, T., Zen, H., Yamagishi, J. and Oura, K., "Speech Synthesis Based on Hidden Markov Models", Proceedings of the IEEE, Vol. 101, No. 5, May 2013

[5] Masuko, T., Tokuda, K. and Kobayashi, T., "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in Proc. Int. Conf. Spoken Lang. Process. (ICSLP), vol. 2, pp. 302–305, Beijing, China, 2000.

[6] Kons, Z. and Aronowitz, H., "Voice transformation-based spoofing of text-dependent speaker verification systems." in Proc. Interspeech, pp. 945-949. Lyon, France, 2013.

[7] Kinnunen, T, Wu, Z., Lee, K.A., Sedlak, F., Chng, E. S. and Li, H., "Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech" in Proc. ICASSP, pp. 4401 – 4404, Kyoto, Japan, 2012.

[8] De Leon, P. L., Stewart, B. and Yamagishi, J., "Synthetic speech discrimination using pitch pattern statistics derived from image analysis" in Proc. Interspeech, Portland, OR, USA, 2012.

[9] De Leon, P. L., Pucher, M. and Yamagishi,J., "Evaluation of the vulnerability of speaker verification to synthetic speech" in Proc. Odyssey IEEE Workshop, pp. 151-158, Brno, Czech Republic, 2010.

[10] Alegre, F., Amehraye, A. and Evans, N., "Spoofing countermeasures to protect automatic speaker verification from voice conversion" in Proc. ICASSP, pp. 3068-3072, Vancouver, Canada, 2013.

[11] De Leon, P. L., Pucher, M., Yamagishi, J., Hernaez, I. and Saratxaga, I., "Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech", IEEE Transactions on Audio, Speech, and Language Processing, vol.20, no.8, pp.2280-2290, 2012.

[12] Wu, Z., Chng, E. S., Li, H., "Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition" in Proc. Interspeech, Portland, OR, USA, 2012.

[13] Wu, Z., Xiao, X., Chng, E. S., Li, H., "Synthetic Speech Detection Using Temporal Modulation Feature" in Proc. ICASSP, pp. 7234-7238, Vancouver, Canada, 2013.

[14] Imai, S., "Cepstral analysis synthesis on the mel frequency scale", in Proc. ICASSP, pp. 93-96, Boston, MA, USA, 1983.

[15] Saratxaga, I., Hernaez, I., Erro, D., Navas, E. and Sanchez, J., "Simple representation of signal phase for harmonic speech models," Electronics Letters , vol.45, no.7, pp. 381,383, 2009.

[16] De Leon, P. L., Hernaez, I., Saratxaga, I., Pucher, M., and Yamagishi, J., "Detection of synthetic speech for the problem of imposture", in Proc. ICASSP, pp. 4844 – 4847, Las Cruces, NM, USA, 2011.

[17] Saratxaga, I. ,Hernáez, I., Odriozola, I., Navas, E., Luengo, I. and Erro, D., "Using Harmonic Phase Information to Improve ASR Rate", in Proc. Interspeech, pp. 1185 - 1188, Makuhari, Japan, 2010.

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

[19] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", in Proc. Eurospeech, pp. 2347–2350, Budapest, Hungary, 1999.

[20] Tokuda, K., Kobayashi, T., Masuko, T. and Imai, S., "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation", in Proc. Int. Conf. Spoken Lang. Process. (ICSLP), vol. 3, pp. 1043-1046, Yokohama, Japan, 1994.

[21] Zen, H., Toda, T., Nakamura, M. and Tokuda, K., "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005", IEICE Trans. Inf. Syst., E90-D(1), pp. 325-333, 2007.

[22] Kawahara, H., Masuda-Kasuse, I. and 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, pp. 187-207, 1999.

[23] 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(2), pp. 184-194, 2014.

[24] Erro, D., Sainz, I., Navas, E. and Hernaez, I., "Improved HNM-based Vocoder for Statistical Synthesizers", in Proc. Interspeech, pp. 1809-1812, Florence, Italy, 2011.

[25] Saratxaga, I., Erro, D., Hernaez, I., Sainz, I. and Navas, E., "Use of harmonic phase information for polarity detection in speech signals", in Proc. Interspeech, pp. 1075–1078, Brighton, UK, 2009.

[26] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering", IEEE Transactions on Audio, Speech, and Language Processing, vol. 19(1), pp. 153-165, 2011.

[27] Maia, R., Akamine, M. and Gales, M.J.F., "Complex cepstrum as phase information in statistical parametric speech synthesis", in Proc. ICASSP, pp. 4581-4584, Kyoto, Japan, 2012.

[28] Drugman, T., Moinet, A., Dutoit, T., & Wilfart, G., Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis. In Proc. ICASSP pp. 3793-3796, Taipei, Taiwan, 2009.

[29] Csapo, T.G.; Nemeth, G. "Modeling Irregular Voice in Statistical Parametric Speech Synthesis With Residual Codebook Based Excitation", Selected Topics in Signal Processing, IEEE Journal of, pp. 209 - 220 Volume: 8, Issue, 2014

[30] Paul, D. B. and Baker, J. M., "The design for the wall street journal-based CSR corpus," in Proc. Workshop on Speech and Natural Language , pp. 357–362, NY, USA, 1992.