INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Waveform generation based on signal reshaping for statistical parametric speech synthesis

Felipe Espic, Cassia Valentini-Botinhao, Zhizheng Wu, Simon King

The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
felipe.espic@ed.ac.uk, cvbotinh@inf.ed.ac.uk, zhizheng.wu@ed.ac.uk, Simon.King@ed.ac.uk

Abstract

We propose a new paradigm of waveform generation for Statistical Parametric Speech Synthesis that is based on neither source-filter separation nor sinusoidal modelling. We suggest that one of the main problems of current vocoding techniques is that they perform an extreme decomposition of the speech signal into source and filter, which is an underlying cause of "buzziness", "musical artifacts", or " muffled sound" in the synthetic speech. The proposed method avoids making unnecessary assumptions and decompositions as far as possible, and uses only the spectral envelope and F0 as parameters. Pre-recorded speech is used as a base signal, which is "reshaped" to match the acoustic specification predicted by the statistical model, without any source-filter decomposition. A detailed description of the method is presented, including implementation details and adjustments. Subjective listening test evaluations of complete DNN-based text-to-speech systems were conducted for two voices: one female and one male. The results show that the proposed method tends to outperform the state-of-the-art standard vocoder STRAIGHT, whilst using fewer acoustic parameters.

Index Terms: speech synthesis, waveform generation, vocoding, statistical parametric speech synthesis

1. Introduction

Statistical Parametric Speech Synthesis (SPSS) has many attractive properties, such as robustness to imperfect data [1] and virtually limitless manipulation of the model's acoustic parameters for speaker adaptation [2], control of emotion [3], style [4], accent, etc. Although hybrid and unit selection-based systems outperform SPSS in terms of naturalness [5], SPSS systems provide higher intelligibility and control.

1.1. Limitations of Statistical Parametric Speech Synthesis

[6] summarizes a widely held view that the lower quality of SPSS, in comparison to waveform concatenation, is due to three problems: over-simplified vocoder techniques that cannot generate detailed speech waveforms, over-smoothing of speech parameters, and acoustic modelling inaccuracy. Other studies have been more formal and have attempted to quantify the relative contributions of these three causes [7, 8, 9]. It seems that about half the degradation is caused by the vocoder alone [10].

1.2. Vocoding Techniques

An SPSS system extracts acoustic parameters from natural speech signals and trains a regression model (Deep Neural Network, Decision Tree, etc.) to predict them from features derived from corresponding text. A vocoder is used to perform tasks that the regression does not generally attempt: acoustic parameter extraction (analysis) & waveform generation (synthesis).

Most vocoders use one of two paradigms: source-filter separation, or sinusoidal modelling. In the former, a source signal that represents glottal pulses or noise produced by turbulent airflow, excites a filter, representing acoustic characteristics of the vocal tract (e.g., STRAIGHT [11, 12], GlottHMM [13, 14], DSM [15, 16]). Sinusoidal models model speech as a sum of sinusoids. The variability of the sinusoids can be modelled by using polynomial functions, adding random noise [17, 18], or randomization of parameters [19], etc. Sinusoidal models are typically not convenient for direct statistical modelling because of the large (and often variable) number of parameters.

2. Motivation

In spite of a proliferation of new vocoders aimed at SPSS in recent years, vocoding remains a significant source of degradation. It appears that the main cause of degradation in source-filter vocoders is the dependence between source and filter [20, 8, 9]: they are in fact not separable. Furthermore, some assumptions lack accuracy. For instance, estimation of filter parameters is made frame-by-frame, assuming that speech production is a linear-time invariant system (LTI) within each frame of analysis. This disregards properties such as the vibration of vocal tract walls, or the abrupt change in the shape of the acoustic cavity at each glottal closure instant (GCI) [21]. These inaccuracies affect the resulting signal which can be perceived as "buzzy", "muffled", with a "phasing" effect, etc.

Although sinusoidal models achieve higher quality than source-filter approaches [22], the decomposition is still suboptimal. Sinusoids cannot accurately represent stochastic components of speech, and the result is "musical artifacts". So, random noise is used to synthesize components over a so-called "maximum voiced frequency" (typically 4kHz) [18]. Other implementations randomize phase [19, for example]. To use sinusoidal vocoders in SPSS, their parameters have to be converted into typical acoustic parameters for SPSS (e.g., spectral envelope, F0, aperiodic energy) which causes degradation.

In summary, we believe that a key problem of current approaches to vocoding is extreme[1] decomposition:

- Many processes of speech production are not well understood, but are approximated by simplistic inaccurate models.
- The dependence between stochastic and deterministic components is hard to capture.
- The vocal tract filter and source signal are not (linearly) separable.

[1] e.g., by attempting to decompose speech into statistically independent source and filter parts.

http://dx.doi.org/10.21437/Interspeech.2016-487

Our proposal is to avoid decomposition, since it is a source of degradation and is not actually necessary to achieve speech synthesis. We should emphasize that the method being proposed here is only for waveform generation; we leave improvements in acoustic parameter extraction for future research. We propose a new paradigm for waveform generation that avoids vocoding, yet is driven by typical acoustic parameters used in typical vocoders, so can be easily used.

3. Proposed Method

The goals for the proposed approach are to:

- Avoid unnecessary extreme decomposition of speech, such as separation into source-filter, stochastic-plus-deterministic, harmonics-plus-noise, etc.
- Focus the design into make a good method for parametric speech synthesis rather than an excellent "speech codec" for copy-synthesis.

There are several poorly understood underlying processes of speech production that are simplistically modelled and/or rely on inaccurate assumptions: the interconnection and dependence between the stochastic and deterministic components of speech; the time-varying and non-linear interaction between the glottal pulses and the vocal tract; the dependence between the phase of the components of speech and other processes involved in the speech production, and so on. Therefore, why not use real speech signals directly in the waveform generation process? By doing so, we might avoid many unnecessary assumptions: the things that we don't understand about natural speech become less important, since they remain intact in this natural speech signal, not separated out in an over-simplified way.

Our goal here is to retain essentially the same regression model that is used in SPSS when driving a vocoder, in order to keep the aforementioned advantages of the statistical parametric approach. We will use a stored natural speech signal (the *base signal*) and "reshape" its characteristics to match the predictions from the regression model; we aim to achieve this with the least possible modification, and in particular without decomposing that stored signal in any way. A complete diagram of a system including the proposed method is shown in Figure 1.

3.1. Acoustic Parameters

We observe that the spectral shape of aperiodic energy is highly correlated with F0, and so it is not necessary to explicitly model or modify it: it is included "for free" in the base signal. Only the spectral envelope and F0 are used as input to the proposed waveform generator: these are the "target" parameters.

3.2. Implementation

The target spectral envelope is derived from the Mel-Cepstrum (MCEP) prediction of the regression model. Whole voiced and unvoiced segments of the utterance (containing several frames, each) are synthesised separately (Sections 3.3 and 3.4), and then concatenated.

3.3. Synthesis of Unvoiced Segments

3.3.1. Database

The database of unvoiced base signals comprises the audio files, spectral envelopes, and spectral envelope averages of just three sustained unvoiced phonemes (/f/, /s/, /ʃ/), recorded by a male speaker in a hemi-anechoic chamber (96kHz; 24 bits).

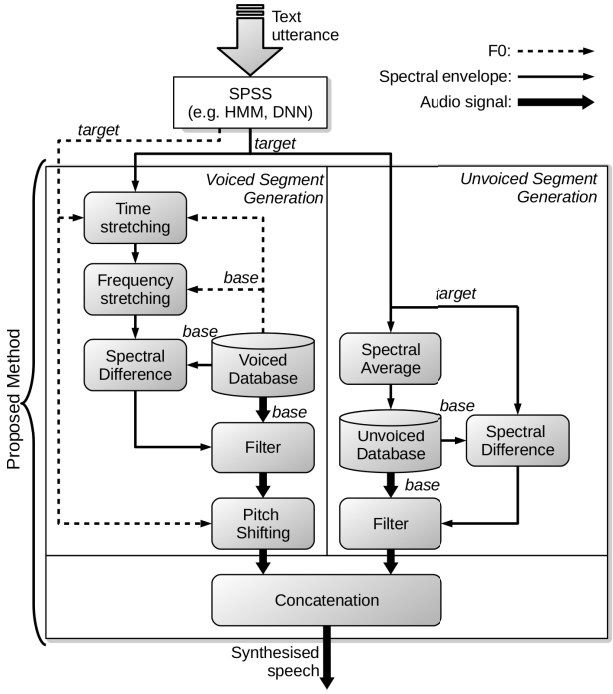

Figure 1: *A SPSS system including the proposed method for waveform generation.*

3.3.2. Spectral Envelope Modification

The spectral envelope of a base signal will be reshaped to match the target. For each target unvoiced segment, one of the three unvoiced base signals in the database is chosen, based on spectral distance to the target: right-hand side of Figure 1. The log spectral envelope difference ([23, 24]) between this and the target is computed, which describes the reshaping needed. Several types of time-varying filters were tried to perform this task (e.g., Finite Impulse Response (FIR), FIR+Overlap-Add (OLA), FIR+Pitch Synchronous Overlap-Add (PSOLA), and MLSA [25]). In informal testing, MLSA was selected.

3.4. Synthesis of Voiced Segments

The synthesis of voiced segments is more complex because they also need to be pitch-adapted. The key design principle is that the processing of base signal waveform is kept to a minimum: filtering, then pitch modification. So, we must construct a time-varying filter that can reshape the base signal's spectral envelope to match the target. The procedure is complicated because the subsequent pitch shifting will change the spectral and temporal structure, so this must be taken into account.

The process is: time-frequency stretching of target spectral envelope, spectral envelope reshaping of base signal, then pitch modification: left-hand side of Figure 1.

3.4.1. Database

The voiced database comprises two audio signals at 96kHz sample rate (higher than the required output sample rate, for reasons that will become clear): the sustained vowel /æ/ uttered by two speakers: a female and a male. We call these the "voiced base signals". The choice of base signal is made per target speaker.

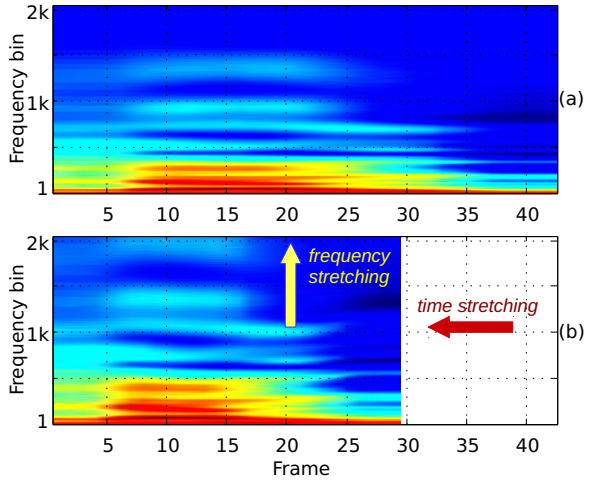

Figure 2: *Example of time-frequency stretching of the target spectral envelope of one voiced segment. (a) Target spectral envelope, from the SPSS regression model. (b) Target spectral envelope stretched to match the base signal's F0. In this example, the target F0 is lower on average than the base signal F0, so the result is that the duration of the target spectral envelope sequence has become shorter (this will be restored in the pitch shifting step, as a side effect), whilst it is stretched in frequency.*

3.4.2. Time-Frequency Stretching of Target Spectral Envelope

The first step is to manipulate the *target* spectral envelope in time and frequency so that its F0 contour matches the F0 contour of the voiced base signal. Later, the final step of processing (Section 3.4.4) will impose the target F0 on the base signal, and as a side effect will change the time/frequency properties of the spectral envelope. We are 'pre-correcting' for that side effect here by moving individual frames of the target spectral envelope sequence closer together or further apart in time, according to the local ratio between base signal F0 and target F0.

Then, the spectral envelope for each frame is stretched (or shrunk) in the frequency direction using cubic spline interpolation, such that the frequency of the first harmonic of the target (if that speech signal were to be created at this point) matches the frequency of the first harmonic of the base signal.

Finally, a uniform frame rate is restored, again using cubic spline interpolation. An example of the complete time-frequency stretching process is shown in Figure 2.

3.4.3. Spectral Envelope Modification

Given the time- and frequency-aligned spectral envelopes of the base signal and the target, we construct a time-varying filter to reshape the base signal to have the target spectral envelope. The filtering is similar to that for unvoiced segments (Section 3.3.2).

3.4.4. Pitch Shifting

The next step is to pitch shift the signal to the target F0 contour. Standard techniques for manipulating F0 independently of spectral envelope / duration (PSOLA, Phase Vocoder, etc.) generate audible artefacts. We avoid such techniques and use simple time-varying resampling to simultaneously impose the target F0 and – as a side-effect – produce exactly the desired spectro-temporal structure. Resampling is performed sample-by-sample using cubic spline interpolation. Since the voiced base signal is sampled at double the required output sample rate, artefacts produced at higher frequencies (aliasing or missing energy) will be removed by downsampling.

This preserves the synchronization, phase relationships, and other dependences between the harmonic and the stochastic components. The natural aperiodicities of the signal are locked to the variations of pitch, as in natural speech (Section 3.1).

Finally, the sequence of voiced and unvoiced segments is concatenated, and downsampled to 48kHz.

3.5. Improvements

Some small improvements are necessary to obtain best results:

- *Spectral Smoothing:* The target spectral envelope derived from MCEPs has reduced resolution at higher frequencies. But the spectral envelope of the base signal is full resolution at all frequencies. Mel-scale smoothing of the base signal's spectral envelope was applied, to make the spectral subtraction (Section 3.3.2) more consistent.
- *Spectral Enhancement:* Spectral envelopes tend to be over-smooth because of the extraction method and/or statistical modelling. To alleviate this, target log spectral envelopes are raised to a power greater than 1 (e.g., 1.1) to enhance peaks.
- *Crossfade:* To avoid artefacts between voiced and unvoiced segments, we crossfade them with 2ms overlap.

4. Experiments

The proposed method is aimed only at improving the naturalness for SPSS systems, so only subjective evaluations are used.

4.1. Subjective Evaluation

Two English text-to-speech voices were built by using a Deep Neural Network-based SPSS system. A female voice based on a speaker called "Laura" was built from 4500, 60 and 67 sentences for training, validation and testing, respectively. A male voice from speaker "Nick" was built using 2400, 70 and 72 sentences. All base signals came from other speakers [2]

MUSHRA-like[3] listening tests were carried out using 30 native English-speaking university students, who each evaluated 30 different sentences (MUSHRA screens) randomly selected from the test sets. For each listener, half of the sentences were the female voice, the rest the male voice. Listeners were asked to evaluate the naturalness of six stimuli (displayed in randomised order) per screen, including four configurations of the proposed method to evaluate the impact of different settings:

- Nat: Natural speech (the hidden reference).
- STR: STRAIGHT.
- SR_all: Signal Reshaping with "ideal" settings: matched-gender voiced base signal, linear-phase filtering, and Mel-scale spectral smoothing (all = all settings ideal)
- SR_gen: as SR_all but base voiced signal is from the opposite gender to target (gen = mismatched gender)
- SR_dp: as SR_all but filtering is not linear phase (dp = distorted phase)
- SR_ns: as SR_all but without Mel-warped spectral smoothing of base signal spectral envelopes (ns = no smoothing)

Listeners were obliged to give one stimulus per screen a score of 100 before proceeding to the next screen.

[2]Durations of base signals: /f/=2.8 secs., /s/=4.4 secs., /ʃ/=2.6 secs., /æ/female=4.6 secs., /æ/male=6.0 secs.

[3]Code available at `http://dx.doi.org/10.7488/ds/1316`

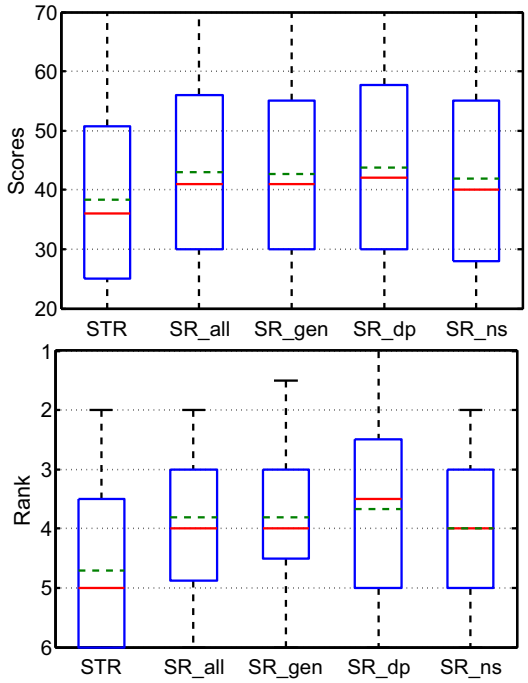

Figure 3: *Results for the female voice. Top: absolute scores; natural speech is omitted (mean score is approx. 100) and the vertical scale is limited to 20–70, for clarity. Bottom: rank (derived from absolute scores within each MUSHRA screen); natural speech is omitted (rank is approx. 1).*

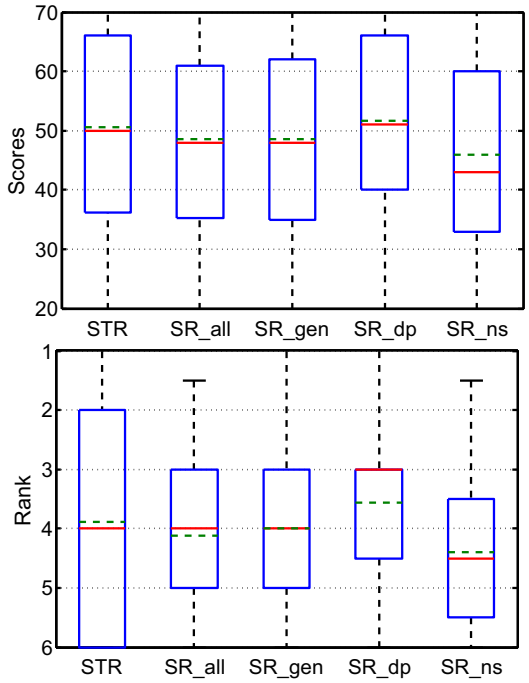

Figure 4: *Results for the male voice, same format as Figure 3.*

4.2. Results

One listener was rejected due to inconsistent scores: natural speech was given a score below 30% several times. Because of the large number of systems to compare, Holm-Bonferroni correction was applied. The Wilcoxon Signed Rank test at $p<0.05$ was used to test statistical significance.

4.2.1. Female Voice

Table 1 and Figure 3 show the results for the female voice. All variants of the proposed method are significantly preferred over STRAIGHT in terms of absolute score. System SR_dp is significantly preferred in terms of both rank and absolute score. SR_dp and SR_all perform significantly better than SR_ns in terms of absolute score; in terms of rank, SR_dp is significantly preferred over SR_ns.

4.2.2. Male Voice

The results of the listening tests for the male voice are shown in Table 1 and Figure 4. SR_dp is significantly better than STRAIGHT with regard to the rank analysis, although there is no significant difference in absolute score. SR_ns is significantly worse than all other systems.

Table 1: *Average MUSHRA score per system in evaluation.*

	system					
	Nat	STR	SR_all	SR_gen	SR_dp	SR_ns
female	99.9	38.3	42.9	42.6	43.7	41.9
male	99.7	50.5	48.5	48.6	51.6	45.9

5. Conclusions, analysis and future work

We have proposed a new paradigm for waveform generation for SPSS that does not decompose waveforms, but instead reshapes a base signal using filtering and pitch manipulation. The test stimuli and response data are available at `http://dx.doi.org/10.7488/ds/1433`.

System SR_dp shows best overall performance: for the female voice, it is significantly better than STRAIGHT in rank and absolute score; for the male voice, it is significantly better than STRAIGHT in rank. We conclude that the proposed method clearly tends to perform better than STRAIGHT. Better relative performance for the female voice could be because:

- It is better to increase, than to decrease, the F0 of the base signal: this moves natural aperiodicities present in the base signal to higher frequencies;
- STRAIGHT is generally worse for female voices than male.

Surprisingly, the distorted phase variant (SR_dp) outperformed the linear phase variant (SR_all); we do not know why.

One advantage of the proposed method is that it needs fewer acoustic parameters than conventional vocoders (only spectral envelope and F0): the SPSS regression model has fewer parameters to predict from the input text features.

Future work includes application of the method to voice conversion, hybrid speech synthesis, join smoothing in concatenation-based systems, and so on.

6. Acknowledgements

The first author is funded by the Chilean National Agency of Technology and Scientific Research (CONICYT). This work was partially supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

7. References

[1] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard challenge 2005," in *Proc. Interspeech*, 2005, pp. 93–96.

[2] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Interspeech*, 2015.

[3] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Monero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit-selection speech synthesis systems applied to emotional speech," *Speech Communication*, vol. 52, no. 5, pp. 394–404, May 2010.

[4] J. Lorenzo-Trueba, R. Barra-Chicote, J. Yamagishi, O. Watts, and J. M. Montero, "Towards speaking style transplantation in speech synthesis," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, Aug. 2013, pp. 179–183.

[5] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Proceedings Blizzard Workshop 2012*, Portland, OR, USA, 2012.

[6] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[7] T. Merritt and S. King, "Investigating the shortcomings of HMM synthesis," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 185–190.

[8] T. Merritt, T. Raitio, and S. King, "Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis," in *Proc. Interspeech*, Singapore, September 2014, pp. 1509–1513.

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

[10] T. Merritt, J. Latorre, and S. King, "Attributing modelling errors in HMM synthesis by stepping gradually from natural to modelled speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Apr. 2015, pp. 4220–4224.

[11] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 34, pp. 187 – 207, 1999.

[12] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*, 1999.

[13] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *Blizzard Challenge 2010 Workshop*, Kyoto, Japan, September 2010.

[14] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. of Interspeech*, Singapore, September 2014, pp. 1969–1973.

[15] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 968–981, March 2012.

[16] T. Drugman and T. Raitio, "Excitation modeling for HMM-based speech synthesis: Breaking down the impact of periodic and aperiodic components," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, Florence, Italy, May 2014, pp. 260–264.

[17] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Fourth European Conference on Speech Communication and Technology, EUROSPEECH 1995, Madrid, Spain, September 18-21, 1995*, 1995.

[18] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 2, pp. 184–194, April 2014.

[19] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP, Journal on Audio, Speech, and Music Processing - Special Issue: Models of Speech - In Search of Better Representations*, vol. 2014, no. 1, p. 38, 2014.

[20] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2733 – 2749, 2008.

[21] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, September 2014.

[22] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 155–160.

[23] K. Kobayashi, T. Toda, and S. Nakamura, "Implementation of f0 transformation for statistical singing voice conversion based on direct waveform modification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5670–5674.

[24] P. L. Tobing, K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Articulatory controllable speech modification based on gaussian mixture models with direct waveform modification using spectrum differential," in *Proc. Interspeech*, Germany, September 2015, pp. 3350–3354.

[25] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, Mar 1992, pp. 137–140 vol.1.