공학석사 학위논문

Modulation Spectrum-based Postfiltering of Synthesized Speech in the Wavelet Domain

파형요소 도메인에서의 변조 스펙트럼 기반 음성합성 후처리

2017년 7월

서울대학교 대학원

전기·정보 공학부

장 세 영

Abstract

This thesis presents a wavelet-domain measure used in postfiltering applications. Quality of HMM-based (hidden Markov model-based) parametric speech synthesis is degraded due to the over-smoothing effect, where the trajectory of generated speech parameters is smoothed out and lacks dynamics. The conventional method uses the modulation spectrum (MS) to quantify the effect of over-smoothing by measuring the spectral tilt in the MS. In order to enhance the performance, a modified version of the MS called the scaled modulation spectrum (SMS), which essentially separates the MS in different bands, is proposed and utilized in postfiltering. The performance of two types of wavelets, the discrete wavelet transform (DWT) and the dual-tree complex wavelet transform (DTCWT), are evaluated. We also extend the SMS into a hidden Markov tree (HMT) model, which represents the interdependencies of the coefficients. Experimental results show that the proposed method performs better.

Keywords: Postfiltering, Modulation spectrum (MS), Discrete wavelet transform (DWT), Dual-tree complex wavelet transform (DTCWT), Hidden Markov tree (HMT)

Student number: 2015-20981

Contents

List of Figures

Chapter 1

Introduction

Speech synthesis has gained popularity over the recent years due to the ever growing demand for natural and convenient interaction between machines and users. Two notable speech synthesis techniques, concatenative speech synthesis and parametric speech synthesis, have been deployed for such purpose. Concatenative speech synthesis, a technique where speech is synthesized by selecting instances of speech directly from the database, has higher quality, but is seriously limited by the size of the data and its intractability; the database has to be very large in order to generate high quality speech, and an entirely new database has to be created in order to adapt to various speaking styles and speaker voices [1]. On the other hand, parametric speech synthesis generates speech parameters from an existing, compact model. Naturally, parametric speech synthesis has gained popularity as sophisticated machine learning techniques emerged, propelling research in novel techniques such as the HMM-based and DNN-based methods [2], [3].

However, speech synthesized from conventional parametric speech synthesis techniques, namely the HMM-based method, have unnatural qualities that is perceived to be artificial by listeners [4]. Although the quantity of the database has a direct effect on the quality of synthesized speech, the primary cause of such degradation is due to the oversmoothing effect, where speech parameter trajectories are smoothed out [4]. Numerous methods have been proposed to alleviate the oversmoothing effect such as employing the global variance or the dynamic features of speech parameters [4], [5].

One measure in observing and analyzing the naturalness of speech is the modulation spectrum (MS). The MS is an effective measure that has been used in various applications such as speaker verification [6] and speech recognition [7]. Studies have shown that the intelligibility of speech is mostly related to the lower modulation frequencies whereas details of speaker dependent characteristics are within the higher modulation frequencies [8], [9].

The degradation of MS in synthesized speech is shown to display a spectral tilt where the modulation spectrum decreases at higher modulation frequency [10]. The conventional postfiltering method mitigates this spectral tilt by directly enhancing the MS through a postfilter which alters the MS of synthesized speech to resemble that of natural speech [10]. Inspired from the MS-based postfiltering technique, numerous other approaches have been proposed which attempt to enhance the modulation spectrum via various techniques such as using line spectral pairs as the input [11] and utilizing DNN-based postfiltering techniques in the MS domain [12] or similarly in the spectral domain [13]. However, there exist problems in directly manipulating the MS, which yields clicking noises when using the MS to perform postfiltering [10].

In this work, the modulation spectrum calculated in the wavelet domain called the scaled modulation spectrum (SMS) is presented and evaluated in subsequent chapters. The SMS is further enhanced using the dual-tree complex wavelet transform (DTCWT) which remedies the drawbacks inherent in discrete wavelet transforms (DWTs). Moreover, an extension of the SMS to a hidden Markov model (HMM) framework is proposed, where individual coefficients of the SMS is modeled using nodes with two states that are connected to form a hidden Markov tree (HMT) structure.

This work is organized as follows. The definition of modulation spectrum and the conventional method is introduced in Chapter 2. In Chapter 3, postfiltering using a simple DWT is described. Chapter 4 details a postfiltering method using the DTCWT, a relatively recent advancement from DWTs. Chapter 5 describes the extension of the studies in Chapter 4 to a HMT framework. Chapter 6 contains the experimental results detailing the comparisons between the conventional and proposed methods. Chapter 7 concludes this work with improvements and future work.

Chapter 2

Modulation Spectrum-based Postfiltering

2.1 Modulation Spectrum

Studies of mammalian auditory systems have shown that they are highly sensitive to modulation of signals due to the ability to perform a multiscale spectrotemporal analysis of acoustic signals [14]. It also accounts for the psychoacoustic aspect of speech such as masking effects. Low modulation frequency corresponds to intelligibility whereas the higher modulation frequency contains speaker characteristics such as the speaker's gender [9].

2.2 Conventional Postfiltering

The MS of speech synthesized from parametric speech synthesis techniques exhibit a spectral tilt due to the oversmoothing effect. This method attempts to remedy this spectral degradation by directly enhancing the MS of synthesized speech to resemble the MS of natural speech. Conventional technique proposed by Takamichi et al. [10] define the MS differently to that from traditional papers defined in acoustics. In [10], speech synthesis applications are defined to be the log-spectral magnitude of the parameter sequence

$$\boldsymbol{s}(\boldsymbol{x}) = [\boldsymbol{s}(1)^\intercal, \cdots, \boldsymbol{s}(d)^\intercal, \cdots, \boldsymbol{s}(D)^\intercal]^\intercal, \tag{2.1}$$

$$\boldsymbol{s}(d) = [s_d(0), \cdots, s_d(f), \cdots, s_d(D_s)]^\intercal, \tag{2.2}$$

$$s_d(f) = \log\Big(\big(\sum_{t=1}^{T} y_t(d)\cos mt\big)^2 + \big(\sum_{t=1}^{T} y_t(d)\sin mt\big)^2\Big) \tag{2.3}$$

where $m = -\pi f/D_s$ and D_s is half of the discrete Fourier transform (DFT) length, y_t denotes the cepstral coefficient at index t, and d and f denote the order of the cepstral coefficient and the modulation frequency, respectively. $\intercal$ denotes the vector transpose. The phase of the DFT of the input signal is preserved and later retrieved in the final synthesis stage.

MS-based postfiltering uses the mean and standard deviation trained from natural speech data to apply the following postfiltering algorithm

$$s'_d(f) = (1-k)s_d(f) + k\Big[\frac{\sigma^N_{d,f}}{\sigma^G_{d,f}}(s_d(f) - \mu^G_{d,f}) + \mu^N_{d,f}\Big] \tag{2.4}$$

where μ and σ denote the mean and the standard deviation of the MS and superscripts N and G denote values from natural speech and the generated speech. The postfiltering coefficient k is a user-controlled parameter in which the degree of postfiltering is applied with a value corresponding to $0 \leq k \leq 1$.

This simple yet effective method greatly enhances the quality of synthesized speech. However, a significant drawback of the MS-based postfiltering technique is the audible clicks that is generated in the filtered sequence [10].

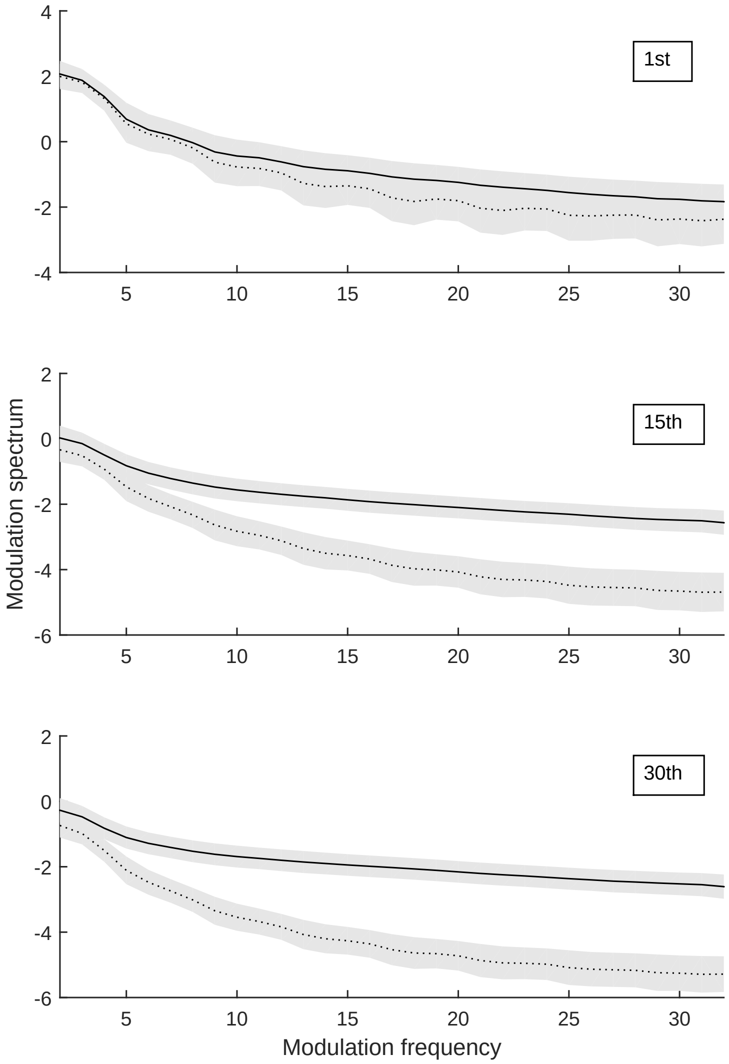


Figure 2.1: Plots of MS of different cepstral orders. The solid line represents the MS of real speech and the dotted line represents the MS of speech synthesized using conventional HMM technique. The gray area represents the standard deviation.

Chapter 3

Discrete Wavelet-based Postfiltering

3.1 Discrete Wavelet Transform

The wavelet transform is an atomic decomposition of a one-dimensional signal by a shifted and scaled versions of a prototype bandpass wavelet function $\psi(t)$ and shifted versions of a lowpass scaling function $\phi(t)$ [15], which is defined as

$$\psi_{j,k}(t) \equiv 2^{-j/2}\psi(2^{-j}t - k) \tag{3.1}$$

$$\phi_{J,k}(t) \equiv 2^{-J/2}\phi(2^{-J}t - k) \tag{3.2}$$

where j denotes the scale factor or the level of decomposition and k denotes the shift factor. Thus, a signal $y(t)$ can be represented as

$$y(t) = \sum_{k} u_k \phi_{J,k}(t) + \sum_{j=-\infty}^{J} \sum_{k} w_{j,k} \psi_{j,k}(t) \tag{3.3}$$

where u_k is the scaling coefficient and $w_{j,k}$ is the wavelet coefficient.

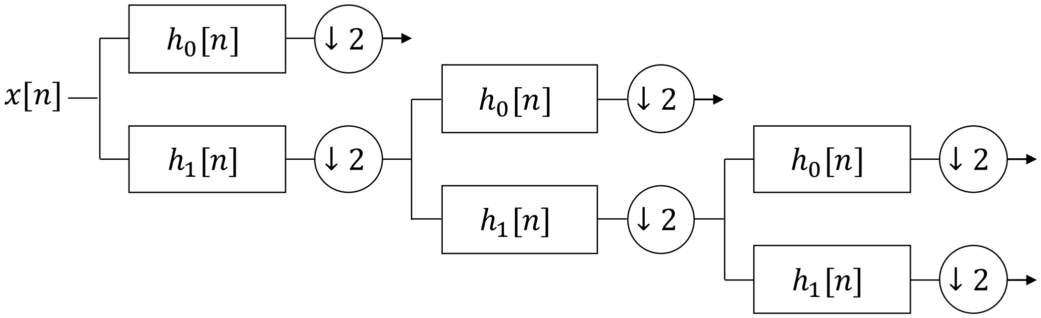


Figure 3.1: Highpass analysis filterbank structure used in the experiment. h_0 and h_1 denote the lowpass and highpass filters, respectively.

While the type of wavelet filter used affects the frequency and temporal resolution for each tile, the filterbank structure determines the tiling in the spectrotemporal domain. Although DWT is flexible, there are severe issues with the DWT; because a sequence is critically sampled, wavelet coefficients near singularities exhibit irregular values. Moreover, filtering and downsampling through a series of nonideal filters results in aliasing, and quantization inherent in digital systems yields artifacts in the reconstructed signal [16]. These issues are addressed in the next chapter where we replace DWT with a better performing counterpart.

3.2 Postfiltering in the Wavelet Domain

Since the MS-based postfiltering method utilizes the Fourier transform, temporal information is not considered in the postfiltering process. Wavelet transforms are more suitable for postfiltering applications because they are localized in both time

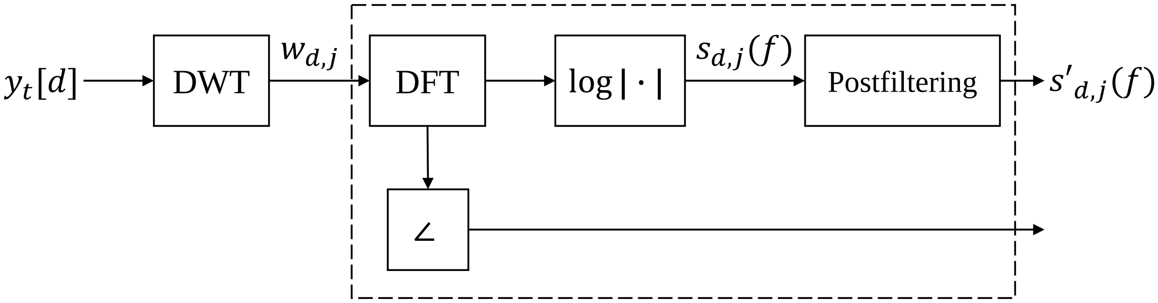


Figure 3.2: Block diagram of postfiltering process using DWT. The dotted square denote the process of calculating the MS.

and frequency. Since high modulation frequency exhibits higher degradation of the modulation spectrum, a highpass analysis filterbank structure shown in Figure 3.1 is used throughout this work. The filter coefficient used for the filterbank is from [17].

To exploit the wavelet transforms for postfiltering applications, we define a new modulation spectrum measure called the scaled modulation spectrum (SMS). This is defined as the MS calculated from the wavelet transform of the speech parameter

$$s_{d,j}(f) = MS\{w_{d,j}(k)\} \tag{3.4}$$

where $MS\{\cdot\}$ denotes the calculation of MS defined in Equation (2.3). Essentially, the SMS shown in Figure 3.3 is the MS calculated at different temporal and spectral resolution which gives a desired property for postfiltering.

The equation for postfiltering is similar to Equation (2.4)

$$s'_d(j,f) = (1-k)s_d(j,f) + k\Big[\frac{\sigma^N_{d,j,f}}{\sigma^G_{d,j,f}}(s_d(j,f) - \mu^G_{d,j,f}) + \mu^N_{d,j,f}\Big] \tag{3.5}$$

where the subscript j indicates the scale of the value. Figure 3.2 shows the overall block diagram of postfiltering in the wavelet domain.

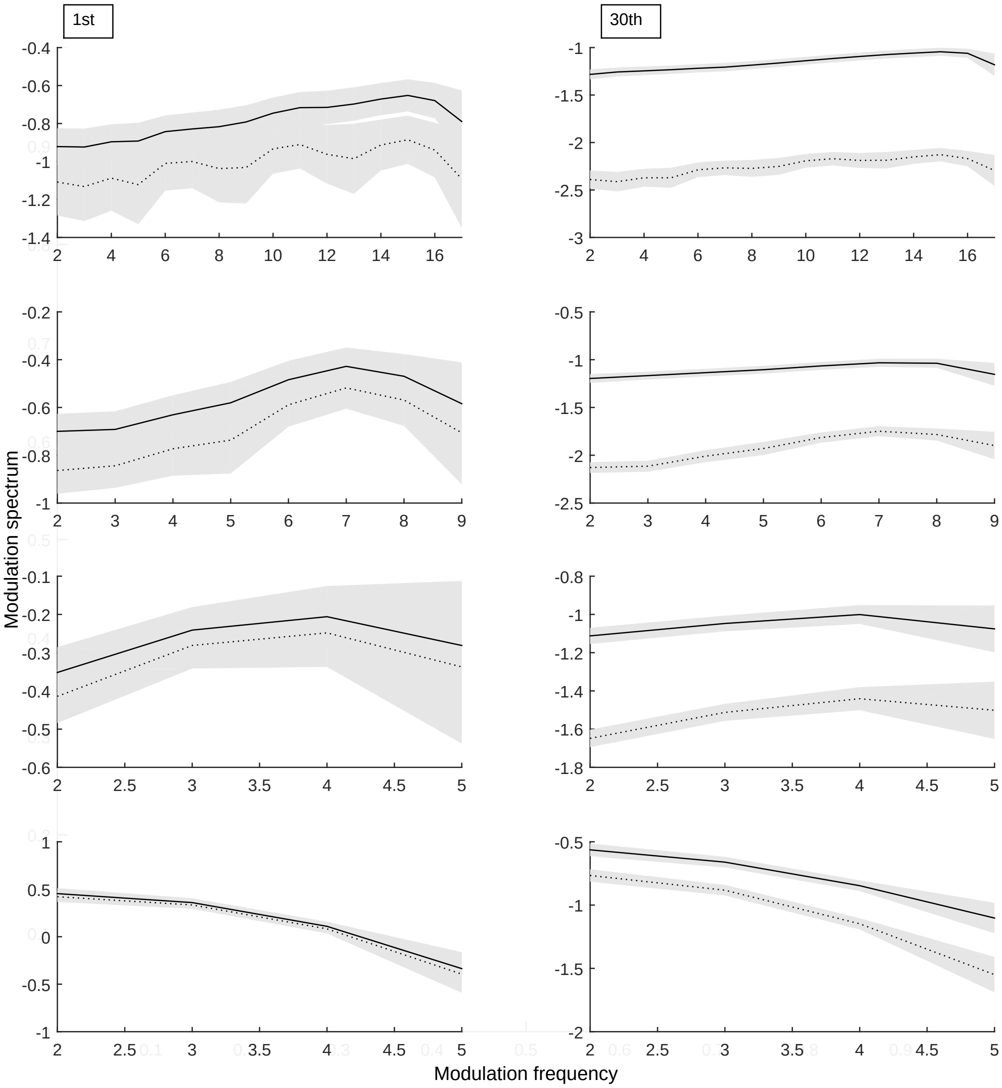


Figure 3.3: Plots of SMS of different 1st and 30th cepstral order. The solid line represents the values of real speech and the dotted line represents the values of synthesized speech. The gray area represents the variance.

Chapter 4

Postfiltering Using Dual-tree Complex Wavelet Transforms

4.1 Dual-tree Complex Wavelet Transform

The DTCWT is a fairly recent enhancement to the standard DWT. The DTCWT overcomes the issues of the DWT mentioned in the previous chapter with a $2N$ redundancy [16]. The wavelet function and the scaling function are represented with complex signals such that

$$\psi_c(t) = \psi_r(t) + j\psi_i(t) \tag{4.1}$$

$$\phi_c(t) = \phi_r(t) + j\phi_i(t) \tag{4.2}$$

where the subscript c, r, and i indicate the complex, real, and imaginary parts. If these two functions form a Hilbert transform pair (90° out of phase) between the real part and the imaginary part, then $\psi_c(t)$ and $\phi_c(t)$ become analytic signals. Projecting these two functions into the standard wavelet decomposition defined in

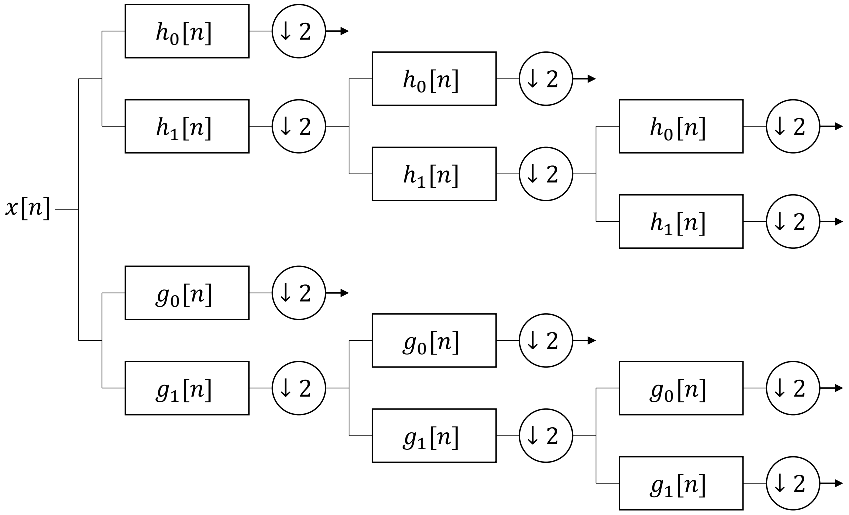


Figure 4.1: Highpass analysis filterbank structure used in the experiment.

Equation (3.3) yields the complex wavelet coefficient

$$w_c(j,k) = w_r(j,k) + jw_i(j,k). \tag{4.3}$$

4.2 Postfiltering Using the DTCWT

SMS using DTCWT is derived using the magnitude of the complex wavelet coefficient, such that

$$s_{d,j}(f) = MS\{|w_{d,j}(k)|\} \tag{4.4}$$

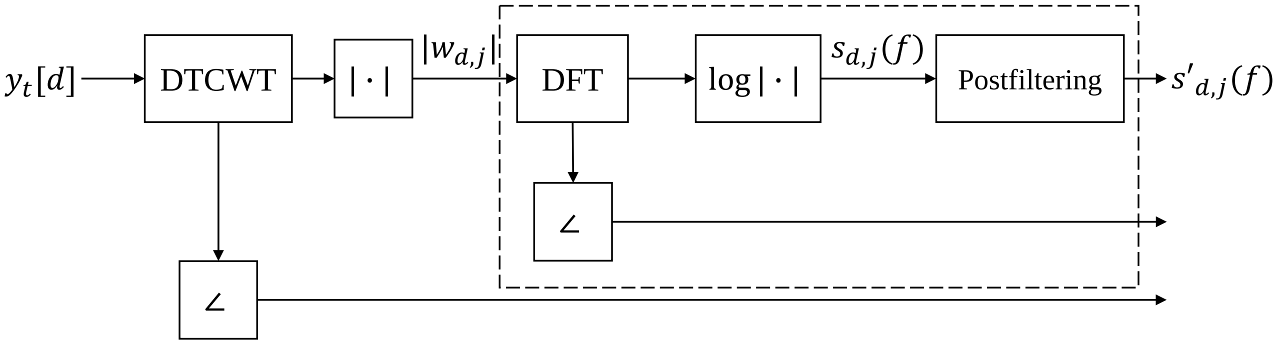


Figure 4.2: Block diagram of postfiltering process using DTCWT. The dotted square denotes the process of calculating the MS.

where $w_{d,j}(k)$ denotes the complex wavelet coefficient of the d-th order at scale j. The phase information, $\angle w_{d,j}(k)$, is stored separately and later used in the synthesis stage. Figure4.2 shows the block diagram of the process.

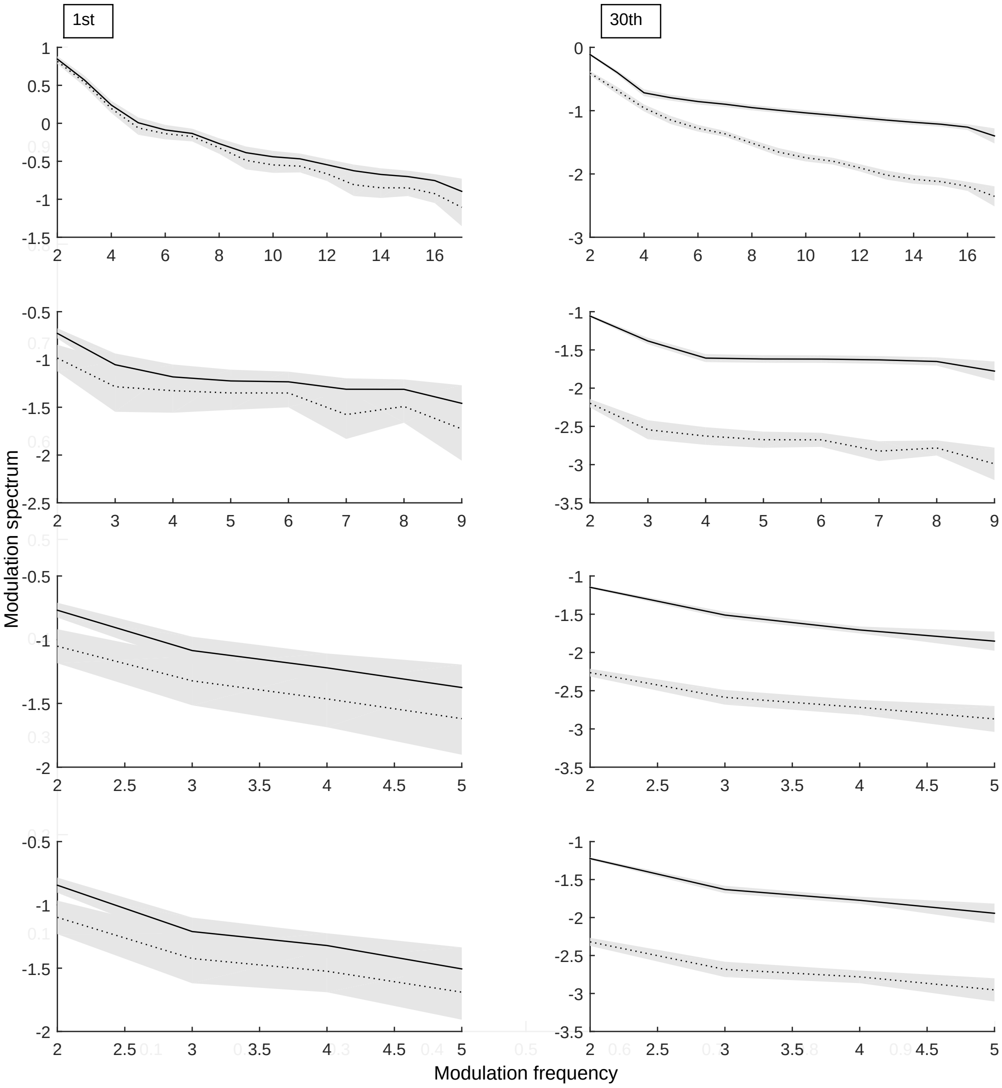


Figure 4.3: Plots of SMS of different 1st and 30th cepstral order. The solid line represents the values of real speech and the dotted line represents the values of synthesized speech. The gray area represents the standard deviation. Rows indicate the level in ascending order.

Chapter 5

Postfiltering Using Hidden Markov Tree Models

5.1 Statistical Signal Processing Using Hidden Markov Trees

Wavelet coefficients of real-world signals exhibit certain interdependencies between their coefficients that can be modeled using different modeling techniques. These interdependencies, namely the clustering and persistence properties of wavelet coefficients, are salient characteristics which can be modeled using a HMM framework [18]. The clustering property states that small or large values of wavelet coefficients are likely to propagate to adjacent coefficients [19], and the persistence property states that small or large values of wavelet coefficients tend to propagate across scales [20], [21]. Clustering property of wavelet coefficients can be modeled using a Markov chain by sequentially linking the states at the same level, whereas the persistence property can be simply modeled by extending the HMM framework

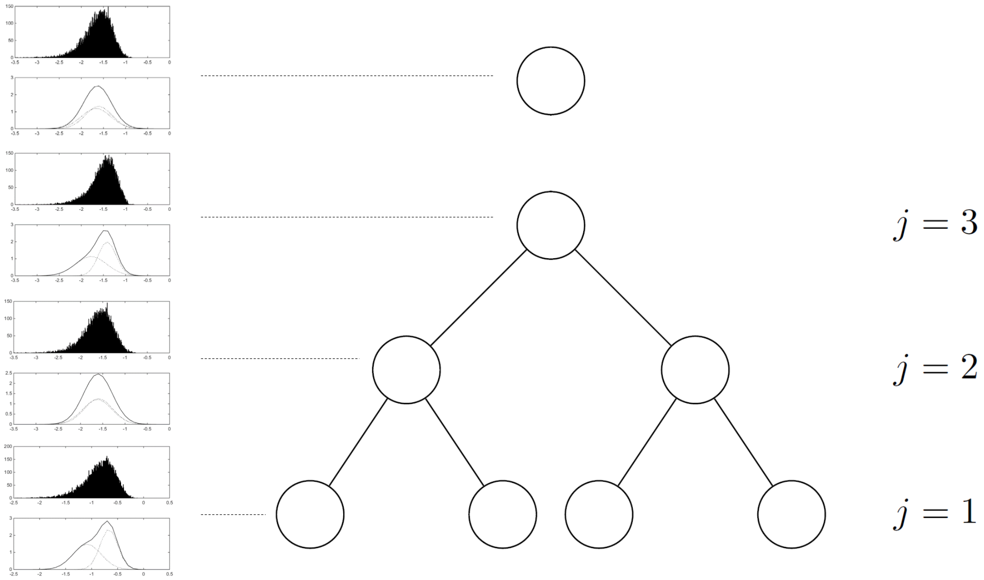
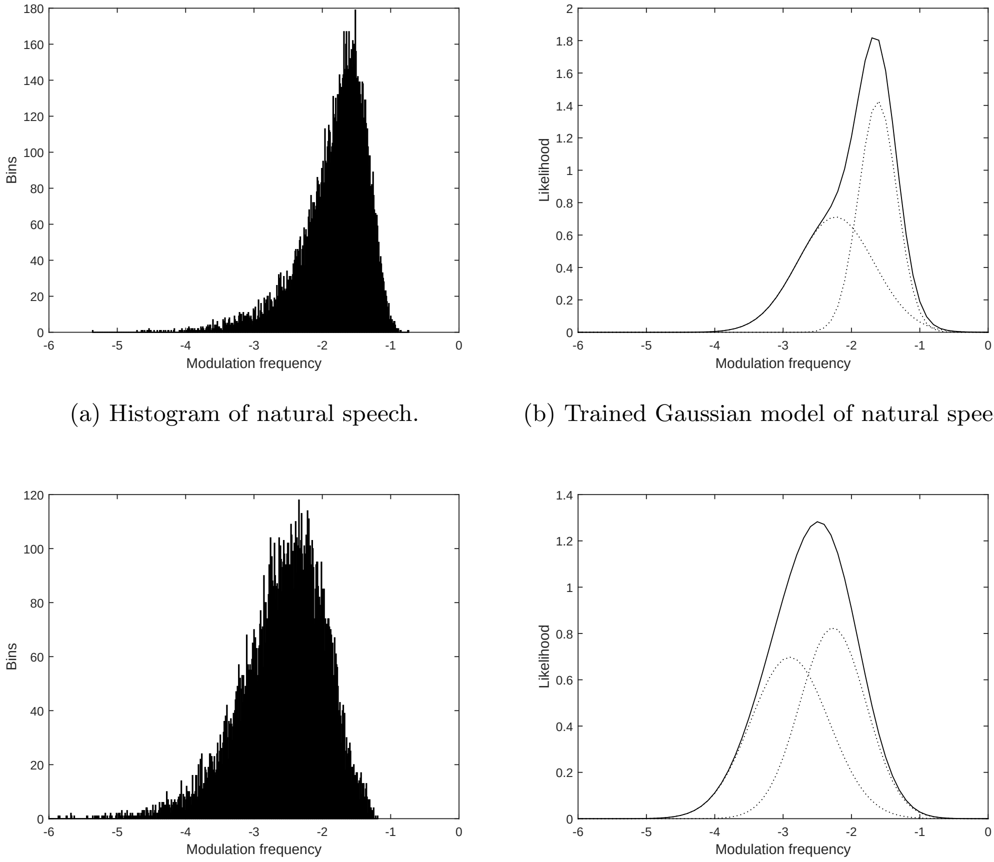


Figure 5.1: A trained HMT model.

into a HMT, where the scaling coefficient is represented with the root node and the wavelet coefficients are represented using subsequent child nodes [18]. Models utilizing these properties have been developed and used in the field of image processing for classification, restoration, and denoising with great success [22], [23].

5.2 Modeling SMS with HMT

By applying the previously mentioned concepts, a HMT model similar to the one described in [23] that exploits the persistence property can be developed for the SMS. This model, depicted in Figure 5.1, contains nodes with states that are classified into "high" and "low" states (denoted as H and L for the 2-state model used in the experiment). The state of child nodes are solely dependent on and are highly likely to inherit the state of the parent node. This is evident in the result of the



(a) Histogram of natural speech.

(b) Trained Gaussian model of natural speech.

(c) Histogram of synthesized speech.

(d) Trained Gaussian model of synthesized speech.

Figure 5.2: Histogram and its corresponding 2-state Gaussian model of a single SMS coefficient calculated from natural speech (above) and synthesized speech (below). $d = 15$, $j = 3$, and $f = 4$.

trained transition matrix being close to the identity matrix. The histogram of SMS coefficients and the trained 2-state Gaussian model are shown in Figure 5, and the resulting Gaussian model closely resembles the distribution of SMS coefficients of natural speech.

Postfiltering in HMT is an extension of Equation (2.4) that accounts for the different states of the HMT nodes

$$s'_d(f) = (1-k)s_d(f) + k\sum_{q\in Q}\frac{\alpha_q}{\sum_{q\in Q}\alpha_q}\Big[\frac{\sigma^{N_q}_{d,f}}{\sigma^{G_q}_{d,f}}(s_d(f) - \mu^{G_q}_{d,f}) + \mu^{N_q}_{d,f}\Big] \tag{5.1}$$

where q and Q denote the state and the number of states, respectively, and α denotes the forward variable. Note that the scale j has been omitted for readability.

The forward algorithm from [24] is used to calculate the state likelihood at each node

$$\alpha_1(q) = \pi_q b_q(s_1) \tag{5.2}$$

$$\alpha_{j+1}(p) = \Big[\sum_{q\in Q}\alpha_j(q)a_{qp}\Big]b_p(s_{j+1}) \tag{5.3}$$

where π indicates the intial state probability distribution, $b(s)$ indicates the observation probability distribution, and a_{qp} denotes the state transition probability from state q to state p. In the experimental setup, each frame of SMS is modeled with four separate HMTs that are concatenated. The overall block diagram of the process is depicted in Figure 5.3.

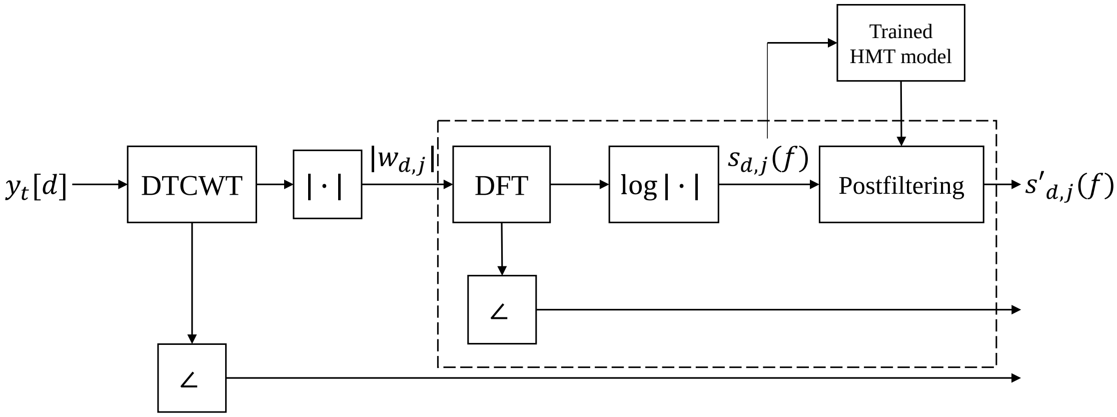


Figure 5.3: Block diagram of postfiltering process using HMT model. The dotted square denote the process of calculating the MS.

Chapter 6

Experimental Results

6.1 Experimental Setup

The frame length and frame shift throughout all of the experiments is set to 32 and 16, respectively. The length of DFT for calculating the MS is set to 64. The length of DFT for calculating the SMS is set to the twice of length of the number of wavelet coefficients in the corresponding scale. The postfiltering coefficient is set to $k = 1$ throughout the experiments. The order of MGC is set to 35. The number of scales for all wavelet transforms is set to $J = 3$. The filter coefficients are from [17] and [25].

Speech samples are retrieved from the CMU ARCTIC US slt set. 593 sentences are used in training. Subjective evaluation is performed on samples randomly selected from 40 synthesized speech. The HMT model is trained using the Contourlet Toolbox with a parameter of 0.0001 for the convergence value, 2 states, and 3 levels.

For the mean opinion score (MOS) test, 10 listeners were asked to score a sample from a range of 1 to 5. 10 samples from each of the 5 different speech samples (HMM,

HMM+MS, HMM+DWT, HMM+DTCWT, HMM+HMT) were tested.

6.2 Results

Figure 6.2 depicts the MS of natural and synthesized speech, as well as the results from different postfiltering methods. The difference between the MS of natural and synthesized speech as well as the MS of the postfiltered speech samples indicate that higher MS, as achieved by the conventional method, does not necessarily mean that the MS is closer to that of natural speech. Hence, subjective measurement is used to evaluate the enhancement in quality. The results of MOS test indicate that postfiltering in the wavelet domain improves the quality synthesized speech. Moreover, although the mean score of HMT is lower than the conventional MS-based method, the standard deviation is significantly higher, indicating that it is inclined to the preference of the listener.

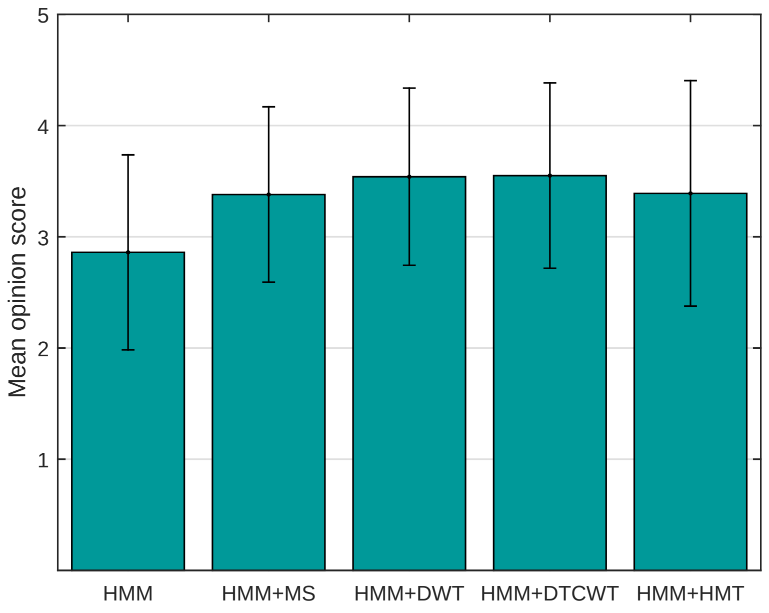


Figure 6.1: Mean opinion score of the different samples.

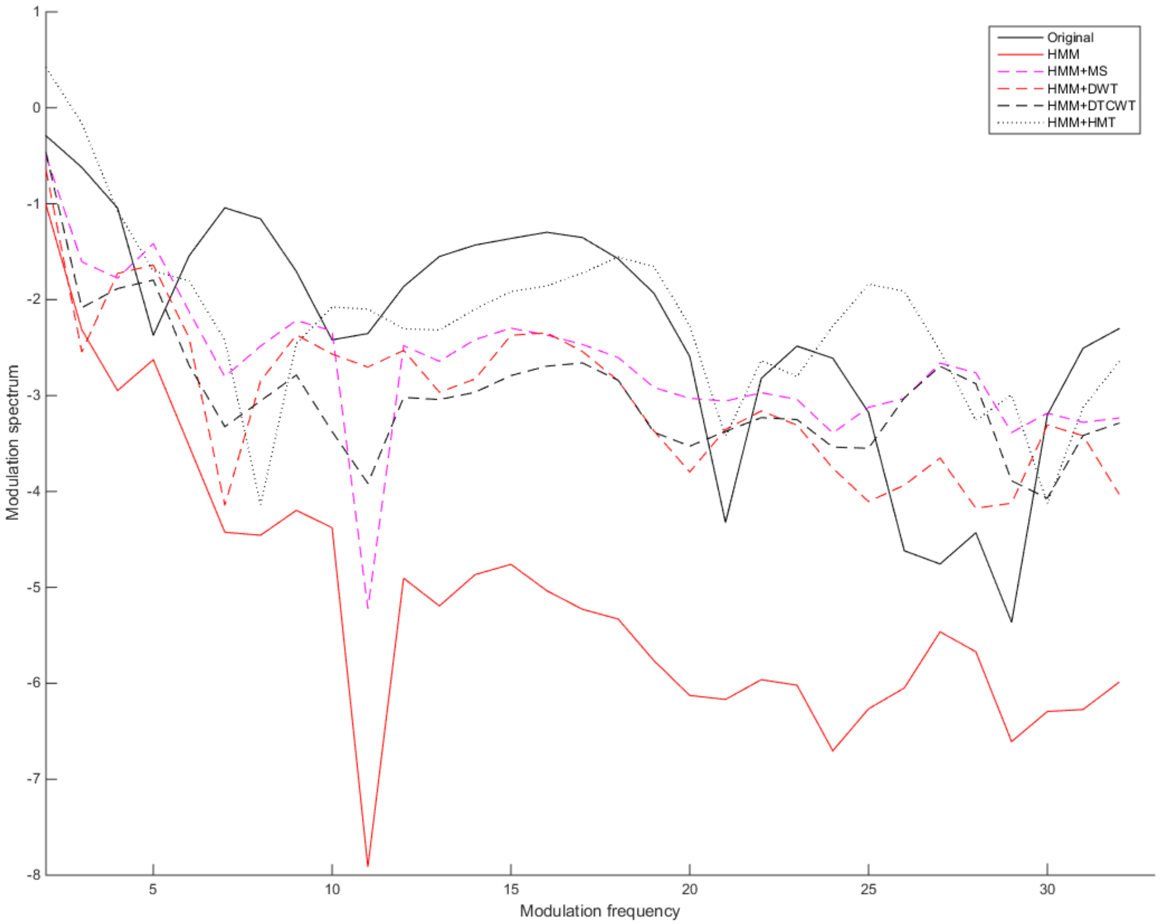
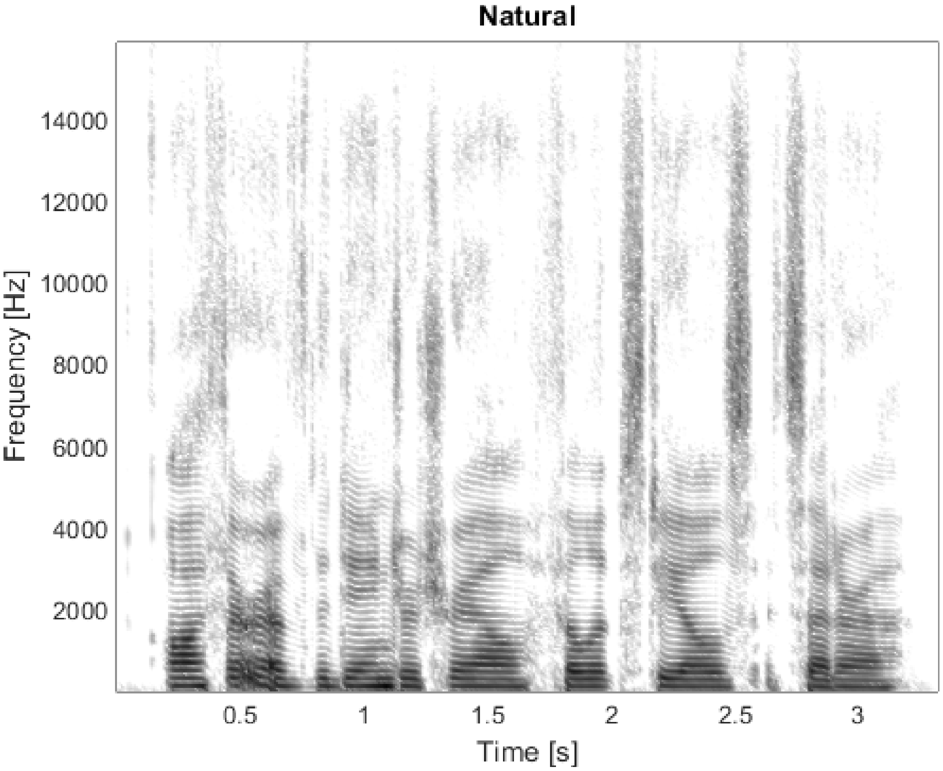
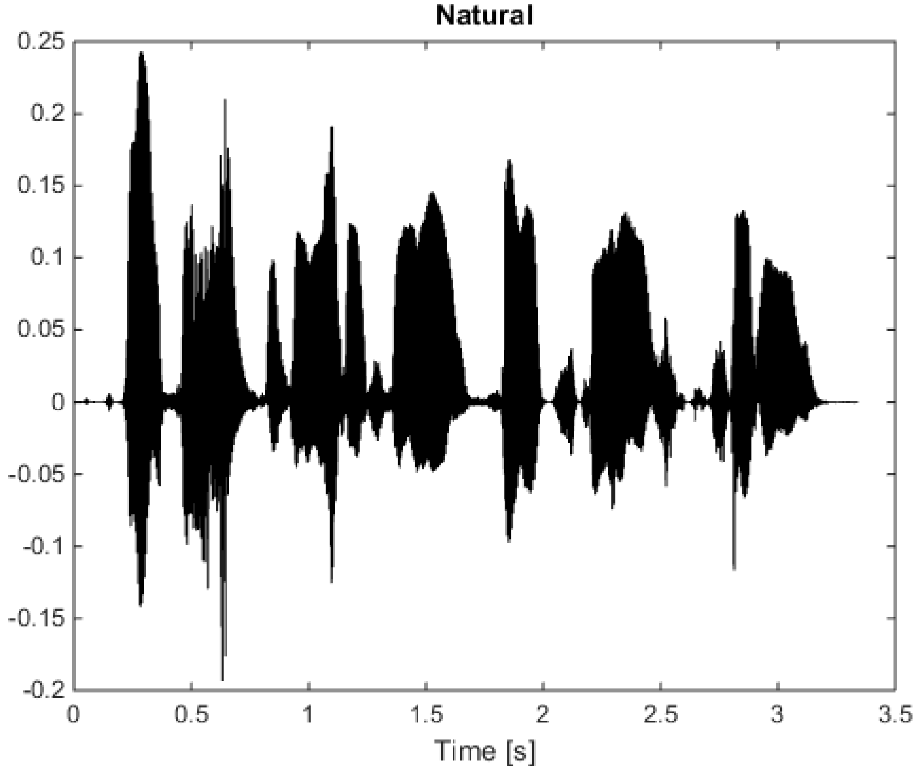


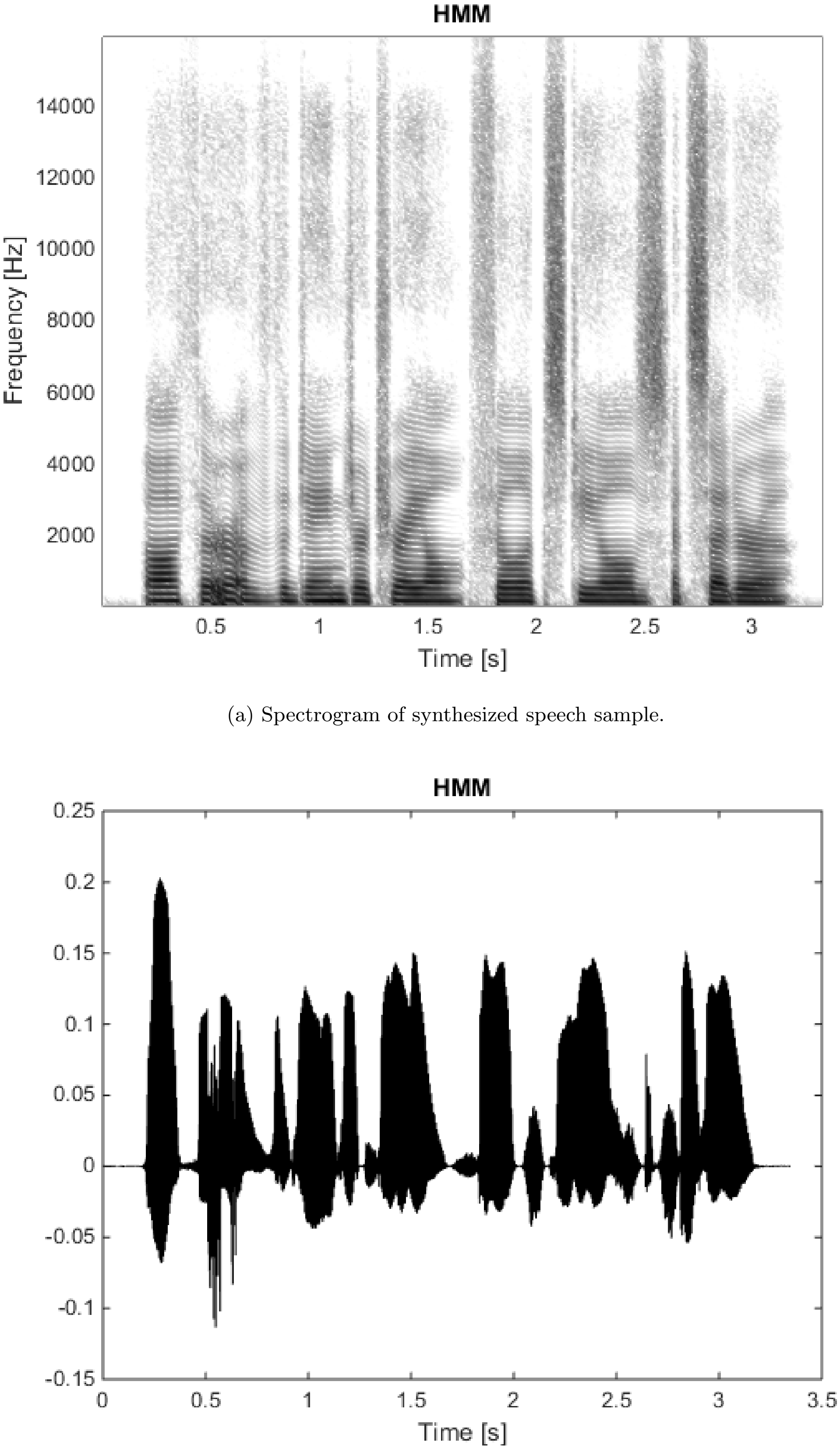
Figure 6.2: MS calculated from different methods. Order of 30. Results indicate that higher modulation spectrum does not necessarily yield better quality.



(a) Spectrogram of natural speech sample.

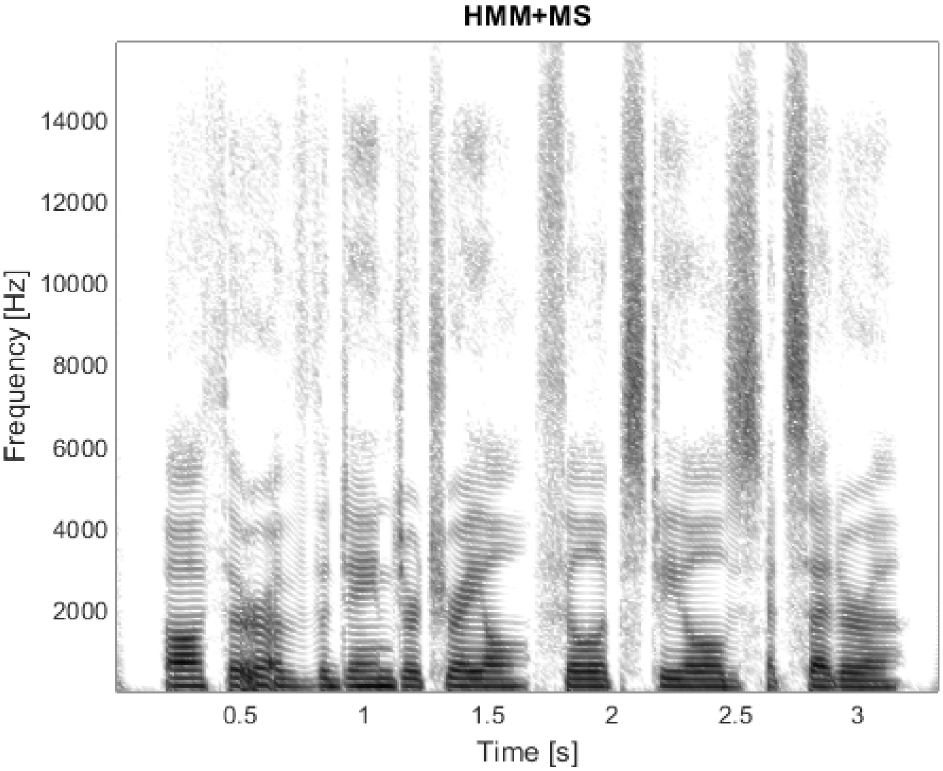


(b) Waveform of natural speech sample.

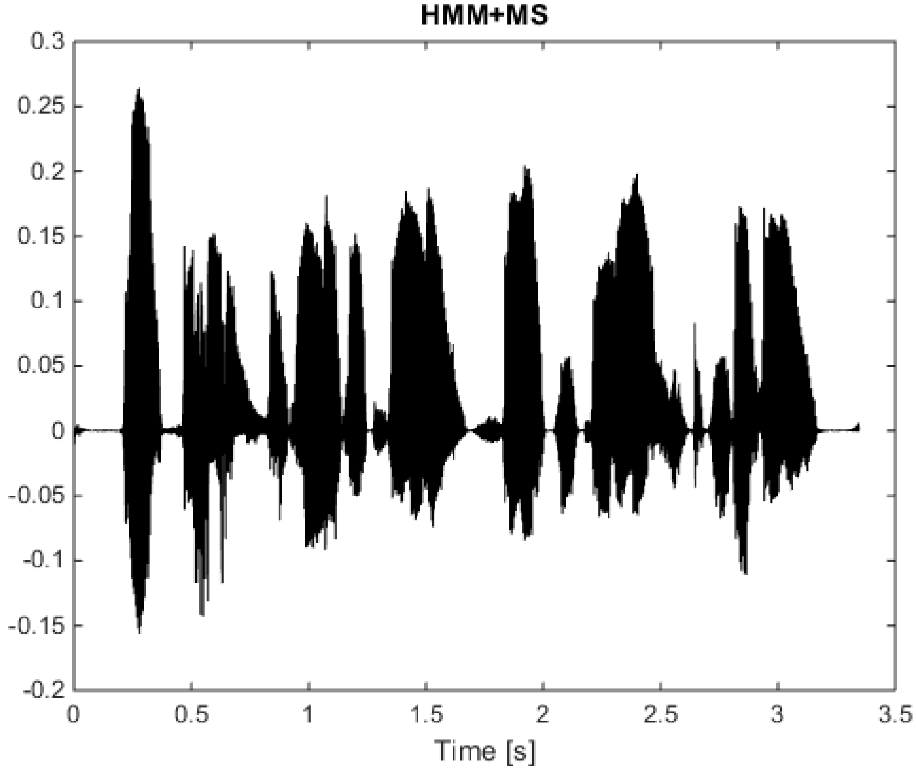


(a) Spectrogram of synthesized speech sample.

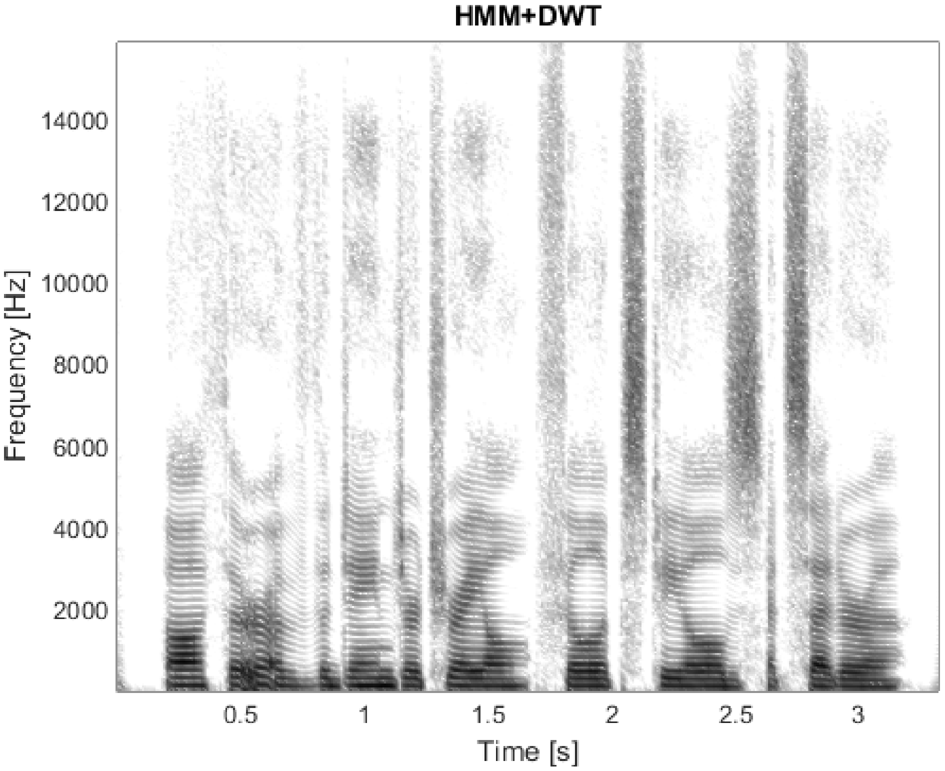
(b) Waveform of synthesized speech sample.



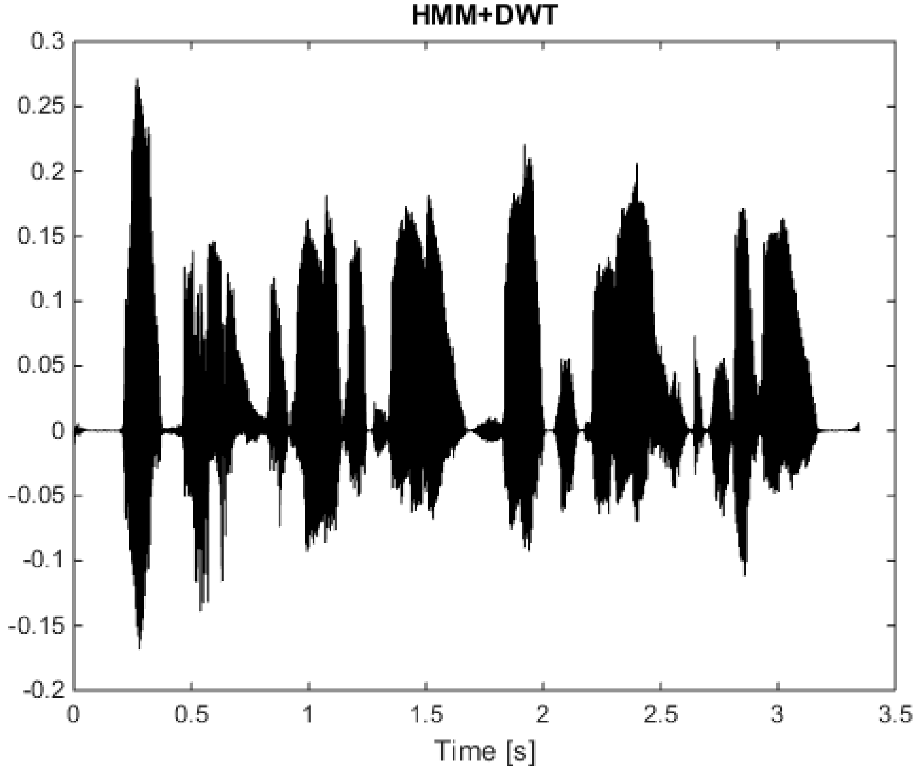
(a) Spectrogram of HMM+MS.



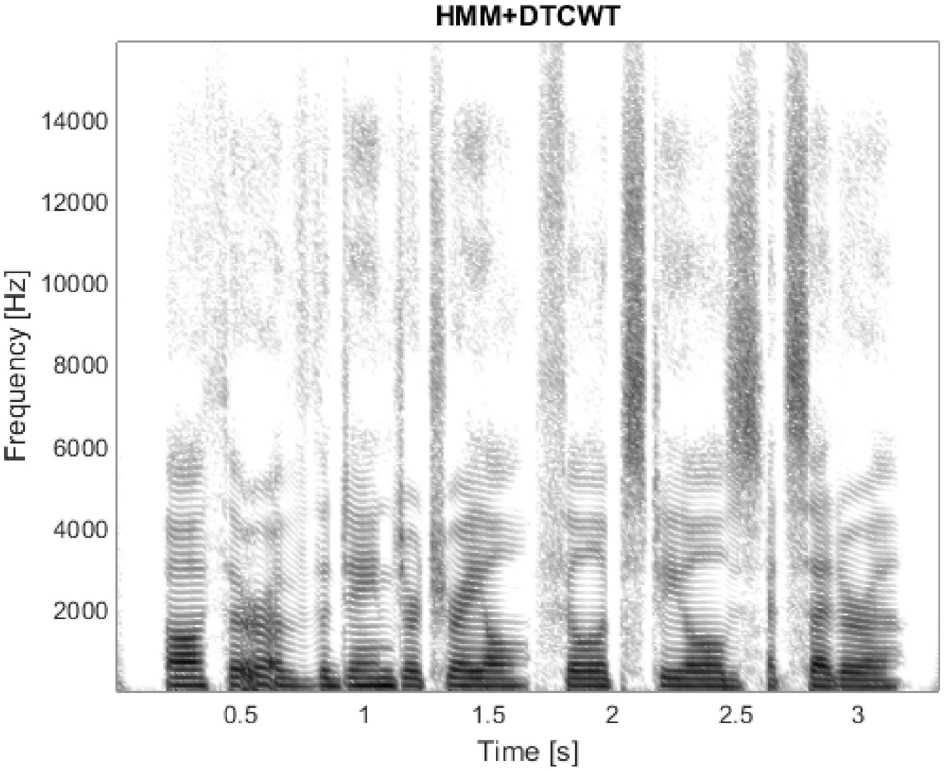
(b) Waveform of HMM+MS.



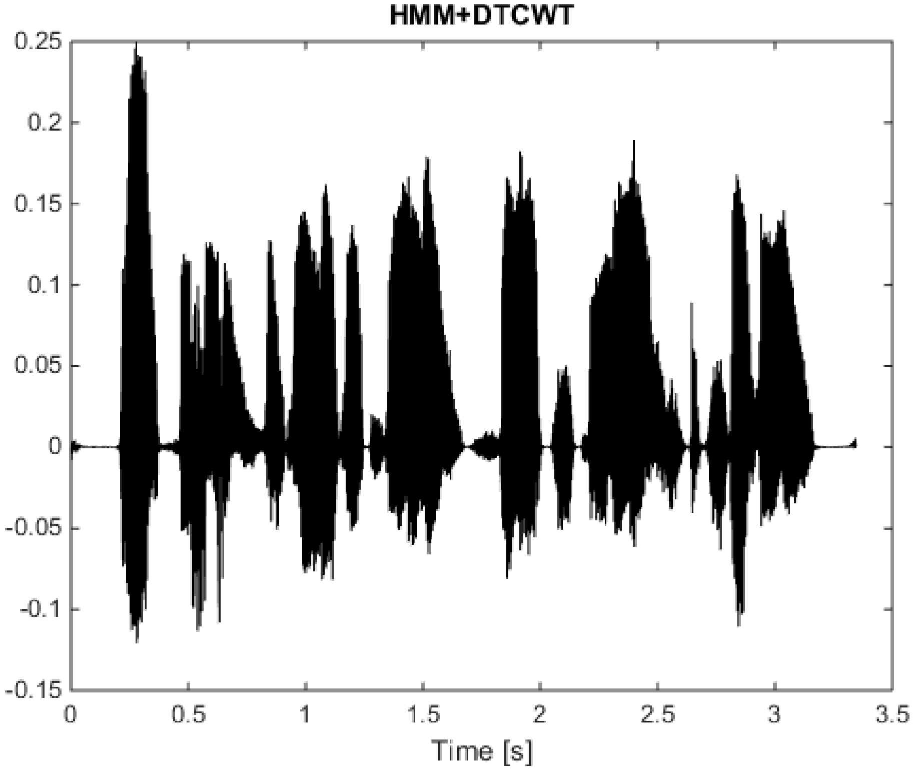
(a) Spectrogram of HMM+DWT.



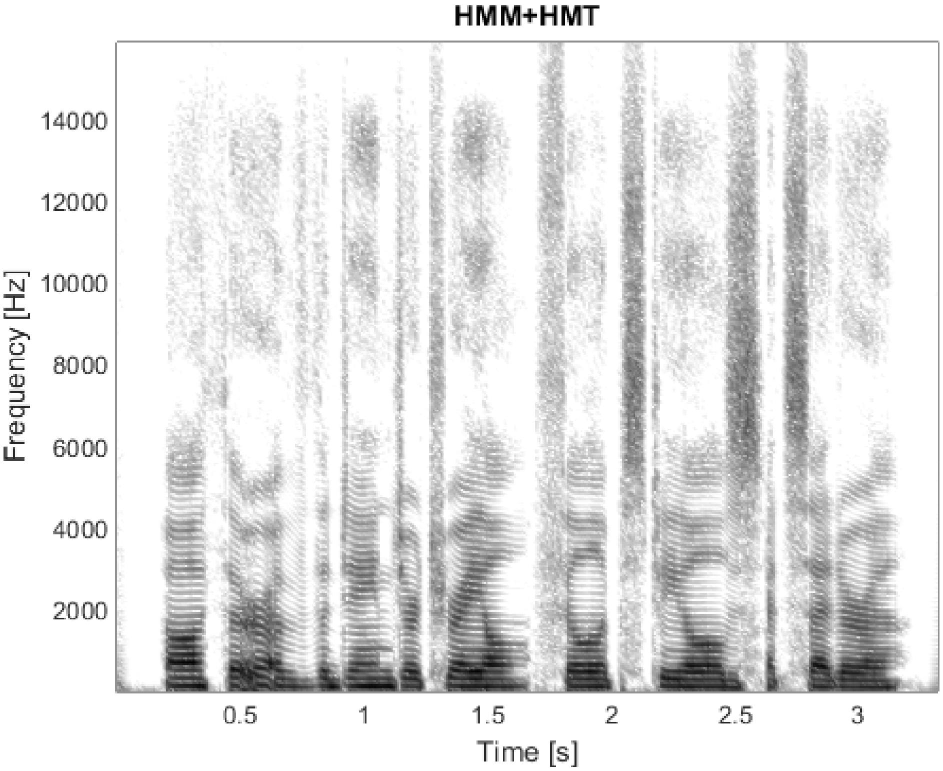
(b) Waveform of HMM+DWT.



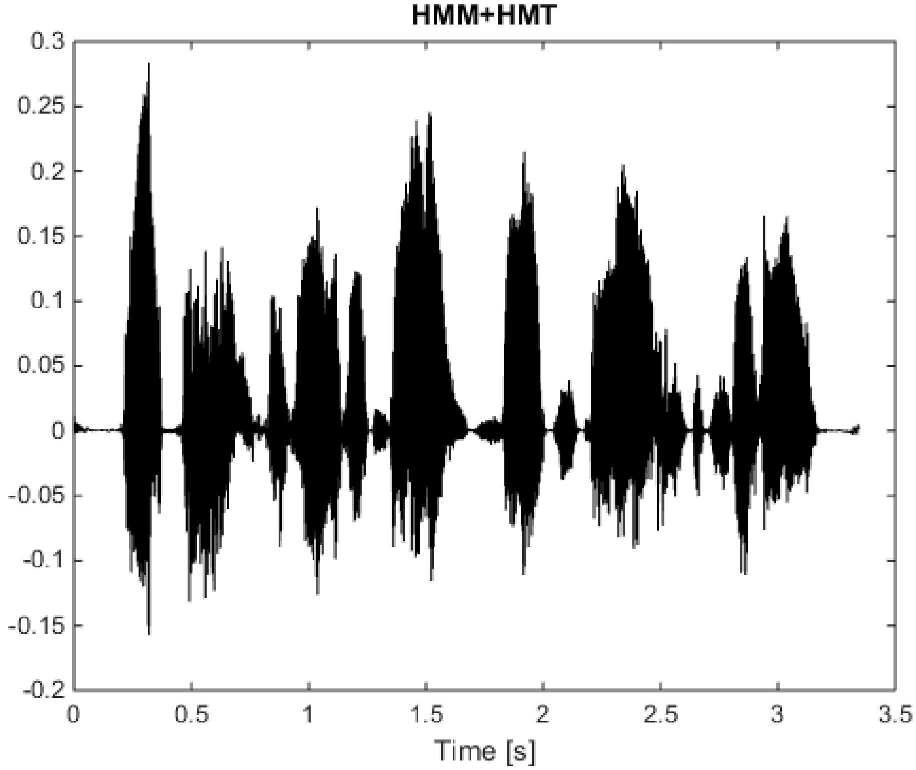
(a) Spectrogram of HMM+DTCWT.



(b) Waveform of HMM+DTCWT.



(a) Spectrogram of HMM+HMT.



(b) Waveform of HMM+HMT.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this work, an improvement to the conventional MS-based postfiltering technique by processing in the wavelet domain was proposed. Because there exists inherent limitations for the conventional framework, namely the issue with spectral resolution in higher frequencies, postfiltering in the wavelet domain has shown to perform better. SMS proved to be a reliable representation of modulation, and is a small step in the direction for further improvements.

Since the MS is not an absolute measure of the naturalness and quality of synthesized speech, high values of MS does not necessarily mean that the postfiltering is more successful. Results from the MOS test corroborate this observation, which shows that listeners prefer postfiltering performed in the wavelet domain.

7.2 Future Work

Further improvements can be made for the current HMT scheme. Namely, the clustering property of wavelet coefficients can be accounted for by linking the nodes at the same level across the trees to create a hidden Markov chain model. Incorporating context dependency can also improve performance, since it yields additional models for specific cases [10]. However, part of this step has to be performed in the synthesis stage which is out of the scope of this study.

Additional transforms can be tested for better performance. The modulated complex lapped transform can be used to calculate a similar form of SMS with a single transformation stage. Additionally, DTCWT with other filterbank structures can be explored.

Finally, employing neural networks to model the SMS coefficients can improve the performance, since DNNs can inherently model both the persistence and clustering properties of the SMS coefficients and other nonlinear properties.

Bibliography

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP*, 1996.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communications*, vol. 51, pp. 1039–1064, Nov. 2009.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, May 2013, pp. 7962–7966.

[4] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[5] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for hmm-based speech synthesis," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 2801–2804.

[6] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 7234–7238.

[7] S. Thomas, S. Ganapathy, and H. Hermansky, "Phoneme recognition using spectral envelop and modulation frequency features," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 4453–4456.

[8] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech perception," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 95, pp. 1053–1064, 1994.

[9] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comput. Biol.*, vol. 5, no. 3, Mar. 2009.

[10] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 755–767, Apr. 2016.

[11] Z. H. Ling, X. H. Sun, L. R. Dai, and Y. Hu, "Modulation spectrum compensation for hmm-based speech synthesis using line spectral pairs," in *Proc. ICASSP*, 2016, pp. 5595–5599.

[12] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, J. Yamagishi, and Z.-H. Ling, "Dnn-based stochastic postfilter for hmm-based speech synthesis," in *Proc. Interspeech*, Sep. 2014, pp. 1954–1958.

[13] L.-H. Chen, T. Raitio, C. Valentini-Botinhao, Z.-H. Ling, and J. Yamagishi, "A deep generative architecture for postfiltering in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2003–2014, Nov. 2015.

[14] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Advances in Signal Processing*, pp. 668–675, 2003.

[15] S. Mallat, *A wavelet tour of signal processing*, 2nd ed. Academic Press, 1999.

[16] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, pp. 123–151, Nov. 2005.

[17] A. F. Abdelnour and I. W. Selesnick, "Nearly symmetric orthogonal wavelet bases," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, May 2001.

[18] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[19] M. T. Orchard and K. Ramchandran, "An investigation of wavelet-based image coding using an entropy-constrained quantization framework," in *"Proc. Data Compression Conf."*, Snowbird, UT, 1994, pp. 341–350.

[20] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 710–732, Jul. 1992.

[21] S. Mallat and W. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inform. Theory*, vol. 38, pp. 617–643, 1992.

[22] G. Zhang and N. Kingsbury, "Variational bayesian image restoration with group-sparse modeling of wavelet coefficients," *Digital Signal Process.*, vol. 47, pp. 157–168, Dec. 2015.

[23] J. Romberg, H. Choi, and R. Baraniuk, “Bayesian tree-structured image modeling using wavelet-domain hidden markov models,” *Signal Processing, IEEE Transactions on*, vol. 10, no. 7, pp. 1056–1068, Jul. 2001.

[24] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[25] N. G. Kingsbury, “Design of q-shift complex wavelets for image processing using frequency domain energy minimisation,” *Proc. IEEE Conf. on Image Processing*, Sep. 2003.