Pitch Prediction from Mel-generalized Cepstrum — a Computationally Efficient Pitch Modeling Approach for Speech Synthesis

Achuth Rao MV, Prasanta Kumar Ghosh
Electrical Engineering
Indian Institute of Science, Bangalore 560012, India
{achuthraomv, prasantg}@ee.iisc.ernet.in

***Abstract*—Text-to-speech (TTS) systems are often used as part of the user interface in wearable devices. Due to limited memory and computational/battery power in wearable devices, it could be useful to have a TTS system which requires less memory and is less computationally intensive. Conventional speech synthesis systems has separate modeling for pitch (F0-model) and spectral representation, namely Mel generalized coefficients (MGC) (MGC-model). In this paper we estimate pitch from the MGC estimated using MGC-model instead of having a separate F0-model. Pitch is obtained from the estimated MGC using a statistical mapping through Gaussian mixture model (GMM). Experiments using CMU-ARCTIC database demonstrate that the proposed GMM based F0-model, even with a single mixture, results in no significant loss in the naturalness of the synthesized speech while the proposed F0-model, in addition to reducing computational complexity, results in ∼93% reduction in the number of parameters compared to that of the F0-model.**

I. Introduction

Wearable devices often use speech-based user-friendly interfaces that utilize text-to-speech (TTS) synthesis units as opposed to text or graphic-based outputs. These devices typically have a limited memory space and computation/battery power [1], [2]. In this paper we propose a pitch modeling approach for the hidden Markov model (HMM) based TTS system (HTS) [3] that reduces both memory requirements and computation complexity compared to the existing pitch modeling approach.

In contrast to the traditional unit concatenation speech synthesis approaches [4], [5], statistical parametric speech synthesis has been effective due to its compact and flexible representation of the voice characteristics [6]. Statistical parametric speech synthesis uses source-filter model of speech and assumes that the phonetic information is conveyed by the spectral envelope, fundamental frequency or pitch (F0) and duration of phones. In HTS, the spectral envelope and pitch are modeled separately from the text input in order to capture spectral (HTS_MGC-model) and pitch (HTS_F0-model) characteristics. Each of these models uses its own decision tree to capture the rich context information [6]–[8]. In the synthesis stage, a given text is converted to the phone sequence and based on the phonetic context, the HMMs of the HTS_MGC-model and HTS_F0-model are concatenated to generate the spectral envelope and F0 respectively. These are used in the source-filter model to synthesize speech. Typically the spectral envelope is represented by a 35-dimension Mel Generalized cepstral co-efficients (MGC) [9] vector.

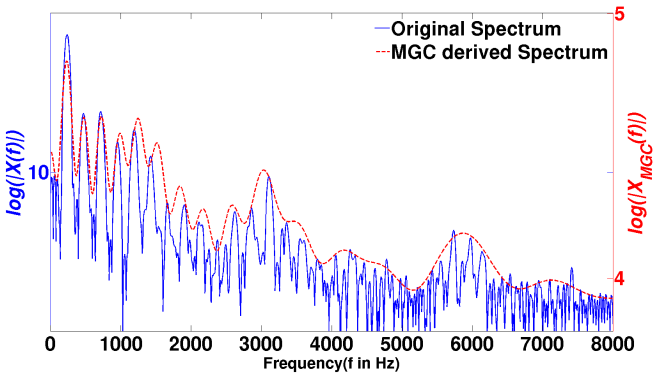

Fig. 1. Illustration of an original magnitude spectrum ($log(|X(f)|)$) and the corresponding MGC derived magnitude spectrum ($log(|X_{MGC}(f)|)$).

In various pitch modification algorithms [10], [11], it has been reported that when the spectral envelope is transformed according to the change in the pitch contour, it results in an improved naturalness of the synthesized speech. Separate modeling of MGC and pitch in HTS may not exploit such relation between the spectrum and the pitch. Instead of a separate modeling for pitch, we explore the statistical relation between the MGC and pitch for modeling the latter in TTS system. The idea of predicting pitch from the MGC is motivated by the work of Syrdal et al. [12] which has demonstrated that localized correlation exists between pitch and spectral envelope or formants, where the correlation between the first formant and the pitch value in case of vowels is investigated. Prediction of pitch has also been explored in the literature from other spectral representations such as MFCC [13], [14], where the pitch is predicted from MFCC using statistical model like GMM to reconstruct speech based on a sinusoidal model of speech. It is known that the harmonic structure in the spectrum of a voiced sound is due to pitch [15] which in turn can be used to estimate F0. However, it is not clear how pitch information could be encoded in 35-dimension MGC. For example, an original magnitude spectrum and the MGC derived magnitude spectrum are shown in Fig. 1. It is clear from Fig. 1 that the

ISBN 978-0-9928626-7-1 © EURASIP 2017

harmonic structure is present in the low frequencies in the MGC-derived spectrum while this fine harmonic structure is smoothed out at higher frequencies. This could be due to the fact that the objective function used to estimate MGCs [9] weighs the error more in the low frequency region than the high frequency one. Based on the harmonic structure in the MGC derived spectrum, we hypothesize that the F0 could be predicted from the MGCs.

The MGCs are calculated so that the unbiased log spectral distance [9] between the generalized log spectrum and the MGC parameterized spectrum is minimum. This process involves a non-linear optimization and there is no closed-form expression representing the MGC given pitch or spectrum. Hence, we explore the GMM based statistical model (MGC_F0-model) to predict the pitch from MGC. During speech synthesis, original MGCs are not available and MGCs are estimated from the HTS_MGC-model as shown in Fig. 2. Therefore, we train the statistical mapping between F0 and MGC estimated from the HTS_MGC-model. Experiments with CMU-ARCTIC corpus [16] reveal that the proposed MGC_F0-model, just with a single mixture GMM, results in a $\sim$93% reduction in the number of parameters required for modeling F0 using HTS_F0-model, without any significant loss in the naturalness of the synthesized speech.

II. Proposed model for estimating pitch from MGC

The proposed MGC_F0-model estimates pitch from the MGC estimated from the HTS_MGC-model in HTS speech synthesizer as shown in Fig. 2. The blocks in blue color in Fig. 2 indicate the components associated with the proposed model. For the HTS_MGC-model, a context-dependent HMM is trained using the MGCs computed from speech and the corresponding text [7] using a decision tree. In order to train the proposed MGC_F0-model, the MGCs are estimated from the trained HTS_MGC-model for the phone labels in the training set with durations same as those of natural speech so that we get time-aligned pitch and MGC. The estimated MGC ($\mathbf{x_i}$) and the original pitch (f_i) in the i-th frame are concatenated to form a single vector $\mathbf{y_i} = [\mathbf{x_i}^T f_i]^T$, where $[.]^T$ is the transpose operator. The Probability density function (PDF) of $\mathbf{y_i}$ is modeled using a GMM with M mixtures.

$$p(f_i, \mathbf{x_i}) = p(\mathbf{y_i}) = \sum_{k=1}^{M} \alpha_k \mathcal{N}(\mathbf{y_i}; \mu_k^y, \Sigma_k^y), \qquad (1)$$

where $\mathcal{N}(\mathbf{y_i}; \mu_k^y, \Sigma_k^y)$ is a normal distribution with the augmented mean vector μ_k^y consisting of the mean MGC vector (μ_k^x) and the mean pitch (μ_k^f), covariance matrix Σ_k^y comprising cross-covariance matrices of the pitch and MGC (Σ_k^{xf} or Σ_k^{fx}), covariance matrix of the MGCs (Σ_k^{xx}) and variance of the pitch (Σ_k^{ff}). Finally, α_k denotes the mixture proportion.

Let there be N voiced frames in the training set. The GMM in eq. 1 is trained with $\{\mathbf{y_i} : 1 \leq i \leq N\}$ using EM algorithm. Once the GMM parameters are learnt, the pitch is estimated from the MGCs using two methods, namely, minimum mean squared error estimation and maximum likelihood estimation with dynamic features. The details of these methods follow.

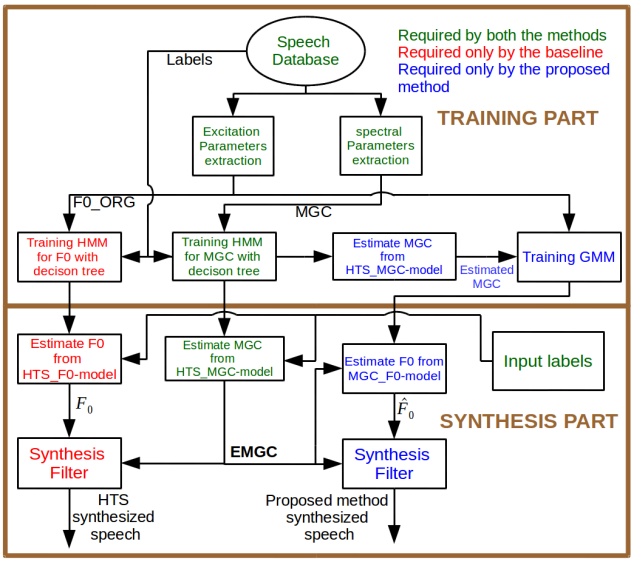

Fig. 2. Block diagram summarizing how the proposed MGC_F0-model is used in the HTS speech synthesizer

A. Minimum mean squared error estimation

The minimum mean squared error estimate minimizes the mean square error between the original pitch f and the estimated pitch $\hat{f}_i$ given the MGCs ($\mathbf{x_i}$) [14]. This leads to $\hat{f}_i = \sum_{k=1}^{M} p_k(\mathbf{x_i}) E_k(\mathbf{x_i})$ where $E_k(\mathbf{x_i}) = (\mu_k^f + \Sigma_k^{fx}(\Sigma_k^{xx})^{-1}(\mathbf{x_i} - \mu_k^x)^T)$ and $p_k(\mathbf{x_i})$ is the weight of k-th mixture given MGC ($\mathbf{x_i}$), i.e.,

$$p_k(\mathbf{x_i}) = \frac{\alpha_k N(\mathbf{x_i}; \mu_k^x, \Sigma_k^{xx})}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x_i}; \mu_j^x, \Sigma_j^{xx})}$$

B. Maximum likelihood estimation using dynamic features

In the case of minimum mean squared error estimator, the pitch is estimated independently in each frame. However, the pitch values in adjacent frames are often correlated. Hence we use maximum likelihood estimation using dynamic features method [17] to exploit this correlation and improve the pitch prediction accuracy by considering the velocity (Δf) and acceleration ($\Delta^2 f$) component of the pitch. The joint PDF of $\mathbf{y_i} = [\mathbf{x_i}^T \; f_i \; \Delta f_i \; \Delta^2 f_i]^T$ is modeled using GMM in eq. 1. The sequence, $f_i (1 \leq i \leq N)$, given the sequence of MGCs $(x_i, 1 \leq i \leq N)$ is estimated by following the work by Toda et al [17].

III. Experiments and Results

A. Corpus description

The experiments are performed using the SLT (female) speaker's data from the CMU-ARCTIC database [16]. The audio is recorded at 16kHz sampling rate and 16-bit resolution. Following the work by [3], the audio is upsampled to 48kHz. The database contains 1132 utterances among which 1000 utterances are selected randomly as the training set and 132 utterances are used for testing [18].

ISBN 978-0-9928626-7-1 © EURASIP 2017

B. Experimental setup:

The MGC and pitch parameters are extracted with a window length of 25ms and a shift of 5 ms. The γ=0 and α=0.55 are used to calculate 35-dimension MGCs. All experiments are performed using HTS [3], [9].

1) Baseline system: HTS_MGC-model is obtained by training context-dependent HMM having 5-states with Gaussian density with diagonal covariance matrix using 105-dimensional MGC vector (including velocity and acceleration coefficients) and the training labels. Similarly, the logarithm of pitch features and its velocity/acceleration coefficients are modeled separately using context-dependent HMM having 5-states with Gaussian density with diagonal covariance matrix to get the HTS_F0-model. This also estimates the voiced and unvoiced labels in each frame and the corresponding model is denoted by HTS_vuv-model. Minimum description length (MDL) based state clustering [8] is performed for both HTS_F0-model and HTS_MGC-model to group the parameters of the context dependent HMMs at the state level. The MDL weighting (λ=1) factor provides a balance between the likelihood improvement and model complexity in MDL. As shown in Fig. 2, given a text during synthesis stage, the MGCs and log(F0) are estimated using HTS_MGC-model and HTS_F0-model considering global variance (GV) to avoid over-smoothing [19], [20]. The duration of each state is found by force aligning the natural speech with the text using the HMM. Given the duration of the each state, the MGCs are estimated from the HMM so that a frame by frame comparison between the synthesized and the natural speech can be made.

2) Proposed method: In the training part of the proposed MGC_F0-model, as shown in Fig. 2, original pitch from all voiced frames is used along with the estimated MGCs from HTS_MGC-model to train a GMM with M mixtures using eq. 1. Given a text during synthesis stage, the MGCs and voiced/unvoiced decisions are estimated from the HTS_MGC-model and HTS_vuv-model. Once the estimated MGCs are available in the voiced frames, the pitch trajectory is estimated using minimum mean squared error and maximum likelihood estimation using dynamic feature methods. These are denoted by MMSE and MLED respectively, where M is varied as 1, 2, 4, 8, 16, and 64.

3) Evaluation: Root mean squared error (RMSE) between the original and the predicted pitch contours in a voiced segment is used as a metric for evaluation. This metric is not highly correlated to the naturalness of synthesized speech. However RMSE has been used to objectively measure the prediction accuracy of the acoustic models [7].

Apart from the pitch prediction error, it is also important to evaluate the quality of the synthesized speech using the proposed MGC_F0-model. For this purpose, we use the Perceptual Evaluation of Speech Quality (PESQ) [21]–[23] of the synthesized speech with respect to the original speech.

In addition to PESQ, subjective listening test is performed with 10 listeners on randomly selected 20 test sentences. The naturalness of the synthesized speech was assessed by the mean opinion score (MOS) test method [24]. Speech samples were presented in random order for all chosen 20 test sentences from MMSE ($M = 1, 8$) and MLED ($M = 1, 8$) as well as using HTS_F0-model. Thus, in the MOS test, each listener provided scores for a total of 100 synthesized audio samples. In the MOS test, after listening to each test sample, the listeners were asked to assign the sample with a naturalness score in a five-point scale – 1-bad, 2-poor, 3-fair, 4-good, 5-excellent [18]. In order to check the consistency of each listener, twenty randomly chosen test samples were repeated in the listening test resulting a total of 120 (=100+20) test samples. All 10 listeners were found to be consistent in providing the naturalness score in at least 16 among 20 repeated test samples.

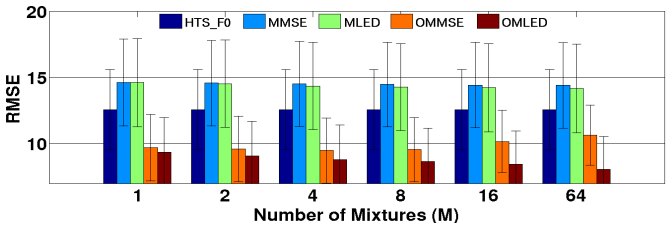

Fig. 3. RMSE based comparison of HTS_F0-model, MGC_F0-models for different values of M with original and estimated MGCs, i.e., MMSE, MLED, OMMSE, OMLED. The RMSE for HTS_F0-model does not depend on M. The bar plot shows the average RMSE with error-bar indicating the SD.

C. Results and discussion

Fig. 3 shows the mean and standard deviation (SD) of the RMSE across 132 test sentences when M is varied. It is clear from the figure that for the proposed method, the average RMSE does not change significantly when M is varied. When averaged across all M, the average RMSE from MMSE and MLED are found to be 2.1Hz and 1.7Hz (absolute) higher than that from the baseline HTS_F0-model.

The MGCs estimated from the HTS_MGC-model are used in the proposed MGC_F0-model. We also examine the accuracy of the pitch predicted using minimum mean squared error and maximum likelihood estimation using dynamic feature methods in the proposed MGC_F0-model when the original MGCs are used, denoted by OMMSE and OMLED. It should be noted that when original MGCs are used, we avoid the error in the pitch prediction contributed by the MGC estimation error from the HTS_MGC-model. Hence using original MGCs would give a lower bound on the RMSE of the predicted pitch using MGC_F0-model. It is clear from Fig. 3 that the average RMSE is lower by $\sim$4Hz when the original MGCs are used compared to when the estimated MGCs are used. In fact, the average RMSE values from OMMSE and OMLED are found to be significantly lower ($p < 10^{-20}$) by 2.74 Hz and 3.86Hz from HTS_F0-model when averaged for all M. This suggests that the pitch could be predicted from the original MGC with lower error than that from the text. From Fig. 3 it is clear that the RMSE of the pitch prediction reduces consistently for OMLED as the number of mixture increases. The average (SD) RMSE for M=64 is 7.73 ($\pm$ 2.38) Hz.

ISBN 978-0-9928626-7-1 © EURASIP 2017

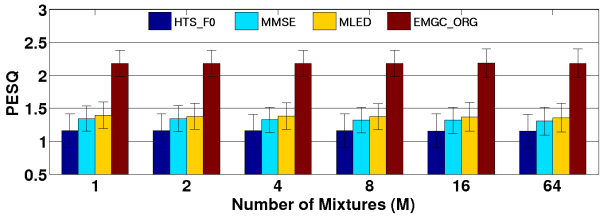

Fig. 4. PESQ based comparison of HTS_F0-model, EMGC_ORG, MMSE and MLED for different M. For HTS_F0-model and EMGC_ORG, PESQ does not depend on M. The bar plot shows the average PESQ with error-bar indicating the SD.

TABLE I
COMPARISON OF THE MOS SCORE FOR DIFFERENT TECHNIQUES.

MOS	HTS_F0-model	MGC_F0-model			
		MMSE		MLED	
		M=1	M=8	M=1	M=8
Average	3.10	2.98	3.05	2.90	3.04
SD	1.33	0.95	1.11	1.15	1.14
p-value	-	0.39	0.68	0.18	0.77

The PESQ is calculated between the synthesized and natural speech for 132 test sentences. PESQ values are shown in Fig. 4. The average PESQ for the proposed method (MGC_F0-model) is higher compared to the baseline system by $\sim$0.16 (absolute) for MMSE method and $\sim$0.21 (absolute) for MLED method when averaged across all M. The difference is statistically significant ($p < 10^{-6}$) for both MMSE and MLED and for each M. We also examine the PESQ between the natural speech and the speech synthesized with estimated MGC and natural pitch contour, denoted by EMGC_ORG. In the absence of any pitch prediction error, this provides an upper bound on the PESQ of the estimated pitch using any pitch prediction model for the given HTS_MGC-model. When averaged across all M, the PESQ for EMGC_ORG is found to be higher than those from HTS_F0-model, MMSE and MLED by 1.02, 0.86 and 0.81 respectively. This indicates that there is room for improvement in PESQ by developing better pitch prediction models.

The MOS scores given by the 10 listeners across 20 utterances for different techniques are shown in Table I. The difference in the MOS scores given by HTS_F0-model and the proposed techniques are not statistically significant as indicated by the p value in Table I indicating that the synthesized speech with the proposed pitch predictive model is as natural as that with the pitch model of the HTS although the later has a lower RMSE in the estimated pitch than the former.

The number of parameters required by the proposed MGC_F0-model in contrast to the HTS_F0-model [18] is shown in Table II. In the proposed MLED, the GMM of M mixtures requires $D \times D \times M + D \times M + M$ number of parameters where D is the dimension of the feature vector. The MLED and MMSE require D=38 (35 MGCs+ static, Δ, Δ^2 pitch) and 36 respectively. The number of parameters required for MLED and MMSE with percentage reduction in comparison to HTS_F0 model is shown in Table II (column titled 'F0'). In column titled 'MGC+F0', it shows the percentage reduction when the number of parameters for HTS_MGC-model are added to F0-models. It is clear that the largest reduction occurs for M=1, for which the naturalness of the synthesized speech does not alter significantly (as seen from Table I) in comparison to the HTS suggesting the benefit of the proposed technique in terms of the number of parameters without compromising the synthesis quality.

TABLE II
COMPARISON OF THE NUMBER OF PARAMETERS BETWEEN HTS_F0-MODEL AND MGC_F0-MODELS FOR DIFFERENT M. PERCENTAGE REDUCTION FROM HTS_F0-MODEL (20K) AND TOTAL NUMBER OF PARAMETERS REQUIRED FOR HTS_F0+HTS_MGC-MODEL (255K) ARE SHOWN IN BRACKETS

M	MMSE		MLED	
	F0	MGC+F0	F0	MGC+F0
1	1.3k(93.4%)	236k(7.4%)	1.4k(92.7%)	237k(7.3%)
2	2.6k(86.8%)	237k(6.9%)	2.9k(85.4%)	238k(6.9%)
4	5.3k(73.7%)	240k(5.8%)	5.9k(70.8%)	240k(5.6%)
8	10.6k(47.5%)	245k(3.7%)	11.8k(41.6%)	245k(3.3%)
16	21.32k(-4.9%)	256k(-0.3%)	23.7k(-16.7%)	256k(-1.3%)

The HTS_F0-model uses maximum likelihood parameter generation [20] and GV method [19] to find the pitch trajectory, which uses the Newton-Raphson method [19], [20], which, in turn, requires computation of the derivative and the Hessian matrix of the likelihood function. This would have a large computational cost, which increases with the duration of the synthesized speech (the number of frames). On the other hand, both MMSE and MLED have closed form solutions which do not require computation of derivative or Hessian. Also both MMSE and MLED methods are applied on the voiced frames unlike the Newton-Raphson based optimization over all frames for speech parameter generation [19], [20]. Thus, the computational cost of MMSE and MLED methods depends on the duration of the voiced region. The length of the voiced region being less than the duration of the entire utterance, the MMSE and MLED methods would have a lower computational cost compared to the HTS_F0-model based pitch trajectory estimation. In fact, the computational cost drops with decreasing M. In particular, the MMSE with M=1 results in a linear transform of MGC to predict the pitch highlighting the computational advantage of the proposed model over HTS_F0-model.

IV. CONCLUSIONS

We propose a pitch modeling approach based on MGC in the speech synthesis framework. The proposed pitch modeling approach reduces the number of parameters as well as computational cost compared to the pitch modeling (HTS_F0-model) present in the HTS speech synthesizer. We show that the pitch prediction accuracy of the proposed model is mainly limited by the quality of the MGCs estimated from the HTS_MGC-model. Even though the pitch prediction accuracy from the estimated MGCs is lower than the HTS_F0-model, the speech synthesized using the proposed pitch model achieves sig-

ISBN 978-0-9928626-7-1 © EURASIP 2017

nificantly higher PESQ score and listening tests show no significant difference in the naturalness of the synthesized speech compared to the HTS_F0-model. The proposed MLED considers exploiting the correlation between the adjacent samples and it also causes over-smoothing. Further improvement on the pitch prediction can be made by considering the GV for MLED similar to the HTS_F0-model but with increased complexity. The proposed MGC based pitch modeling can also be used in Deep Neural Network (DNN) based speech synthesis framework [7].

REFERENCES

[1] Sang-Jin Kim, Jong-Jin Kim, and Minsoo Hahn, "HMM-based korean speech synthesis system for hand-held devices," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 4, pp. 1384–1390, 2006.

[2] Kunjal Parikh, Khaled Ahmed, Naoki Matsumura, David Gottardo, Ramon Cancel, Brian Girvin, and Ronald Woodbeck, "42-3: Invited paper: Requirements for next generation wearable display and battery technologies," in *SID Symposium Digest of Technical Papers*. Wiley Online Library, 2016, vol. 47, pp. 570–573.

[3] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, and Keiichi Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0.," *Sixth ISCA Workshop on Speech Synthesis*, pp. 294–299, 2007.

[4] Alan W Black and Paul Taylor, "CHATR: a generic speech synthesis system," *Proceedings of the 15th conference on Computational linguistics-Volume 2*, pp. 983–986, 1994.

[5] Andrew J Hunt and Alan W Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *International Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 1, pp. 373–376, 1996.

[6] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proc. Eurospeech*, pp. 2347–2350, 1999.

[7] Heiga Zen, Alan Senior, and Martin Schuster, "Statistical parametric speech synthesis using deep neural networks," *International Conference Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7962–7966, 2013.

[8] Koichi Shinoda and Takao Watanabe, "MDL-based context-dependent subword modeling for speech recognition.," *The Journal of the Acoustical Society of Japan (E)*, vol. 21, no. 2, pp. 79–86, 2000.

[9] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation.," *Proc. ICSLP*, 1994.

[10] Kimihito Tanaka and Masanobu Abe, "A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F0," *International Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 2, pp. 951–954, 1997.

[11] Alexander Kain and Yannis Stylianou, "Stochastic modeling of spectral adjustment for high quality pitch modification," *International Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 2, pp. II949–II952, 2000.

[12] Ann K Syrdal and Shirley A Steele, "Vowel F1 as a function of speaker fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 78, no. S1, pp. S56–S56, 1985.

[13] Xu Shao and Ben Milner, "Pitch prediction from MFCC vectors for speech reconstruction," *International Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 1, pp. I–97, 2004.

[14] Ben Milner and Xu Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 24–33, 2007.

[15] James L Flanagan, *Speech analysis synthesis and perception*, vol. 3, Springer Science & Business Media, 2013.

[16] John Kominek and Alan W Black, "The CMU ARCTIC speech databases," *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[17] Toda Tomoki, Alan W Black, and Keiichi Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[18] Kai Yu and Steve Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[19] Toda Tomoki and Keiichi Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[20] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *International Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, vol. 3, pp. 1315–1318, 2000.

[21] Yi Hu and Philipos C Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[22] Milos Cernak and Milan Rusko, "An evaluation of synthetic speech using the pesq measure," in *Proc. European Congress on Acoustics*, 2005, pp. 2725–2728.

[23] J Sangeetha, S Jothilakshmi, S Sindhuja, and V Ramalingam, "Text to speech synthesis system for tamil," *Int J Emerging Tech Adv En*, vol. 3, pp. 170–5, 2013.

[24] Mahesh Viswanathan and Madhubalan Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale," *Computer Speech & Language*, vol. 19, no. 1, pp. 55–83, 2005.

ISBN 978-0-9928626-7-1 © EURASIP 2017