ADAPTIVE FILTERING FOR HIGH QUALITY HMM BASED SPEECH SYNTHESIS

Luis Coelho[1], *Daniela Braga*[2]

[1] Instituto Politécnico do Porto, ESEIG, Porto, Portugal (lcoelho@eu.ipp.pt)
[2] MLDC - Microsoft Language Development Center, Lisbon, Portugal (i-dbraga@microsoft.com)

ABSTRACT

In this work an adaptive filtering scheme based on a dual Discrete Kalman Filtering (DKF) is proposed for Hidden Markov Model (HMM) based speech synthesis quality enhancement. The objective is to improve signal smoothness across HMMs and their related states and to reduce artifacts due to acoustic model's limitations. Both speech and artifacts are modelled by an autoregressive structure which provides an underlying time frame dependency and improves time-frequency resolution. The model parameters are arranged to obtain a combined state-space model and are also used to calculate instantaneous power spectral density estimates. The quality enhancement is performed by a dual discrete Kalman filter that simultaneously gives estimates for the models and the signals. The system's performance has been evaluated using mean opinion score tests and the proposed technique has led to improved results.

Index Terms— Kalman filtering, Spectral Analysis

1. INTRODUCTION

In the last few years HMM based text to speech systems (TTS) gained paramount importance on the speech community. One of the main advantages of this technique when compared with the unit selection and concatenation method is the fact that it is possible to build a good quality TTS with as few as eighty sentences [1] and voice transformation can be easily achieved by model adaptation [2]. During signal generation, if a global variance [3] is not considered, fast parameters variations across model and state transitions may be introduced even if dynamic features are used. A final low-pass filter can by introduced to reduce sudden spectral changes and artifacts but voice quality will always suffer. In [4] a CELP based coding scheme is used with good results but at the cost of additional complexity during synthesis and training.

Without restrictions to the speech coding technique used on the HMM framework we propose the replacement of the above mentioned final low-pass filter by an adaptive Kalman filter assuming an underlying autoregressive (AR) structure for the signal [5]. Additionally the generated speech artifacts, considered independent, are simultaneously modelled and included in the Kalman estimation. The AR structure gives an intrinsic correlation between consecutive time samples which leads to the desired smoothing and provides accurate estimations of the power spectral density (PSD), for both signal and artifacts. For a sufficient model order a high frequency resolution is achieved which can be used as the base for additional time-domain [6] and spectral-domain analysis (spectral-subtraction [7, 8] and signal-subspace embedding [9]). The proposed model is used for adaptive signal smoothing and artifact modelling

The rest of this paper is organized as follows. Section II discusses the used methodology with details on the state-space model for the signal and the kalman filter iterative algorithm. In Section III the results of the conducted experiments are presented and then widely discussed in Section IV. The conclusions are presented in section V.

2. METHODOLOGY

The discrete Kalman filter based tracking approach requires a discrete state-space model with the form:

$$\mathbf{x}(k) = \mathbf{F}\mathbf{x}(k-1) + \mathbf{G}\mathbf{w}(k) \quad (1)$$

$$\mathbf{y}(k) = \mathbf{H}\mathbf{x}(k) + \mathbf{v}(k) \quad (2)$$

where $\mathbf{x}(k)$ is the state vector, $\mathbf{y}(k)$ is the output or measurement vector, **F** is the (state-space) process matrix that relates previous and present states, **G** is the input weight vector, $\mathbf{w}(k)$ is the input vector, **H** is the output matrix and $\mathbf{v}(k)$ is a possible output disturbance. Equation 1 is a general vector Gaussian-Markov model with $\mathbf{w}(k)$ as a driving excitation vector.

An AR recursion has been used for signal representation with $s(k)$ as the speech signal under analysis at instant k, $\{a_i\}_{i=1}^{M}$ are the model parameters, $\{s(k-i)\}_{i=1}^{M}$ are delayed samples of the signal and $w(k)$ is assumed to be the excitation component at time instant k. This can be written using a compatible state vector $\mathbf{x}(k)$ and a process matrix in controllable canonical form. The matrix content vary with time because the AR coefficients a_i must be adjusted as the signal changes. With the variable state transition matrix we have a dynamic vector Gaussian-Markov state-space model.

The input weight matrix **G** that interfaces the driving excitation $\mathbf{w}(k)$ and the process output matrix is:

$$\mathbf{G}^T = \mathbf{H} = (1 \overbrace{0 \ldots 0}^{M-1}) \tag{3}$$

Coherently with equation 3 the output $y(k)$ and $v(k)$ are values and interface with the measurement error, with the last having variance σ_s^2.

The artifacts are also modelled by an AR process with order N that is uncorrelated with the signal. This can be considered a valid assumption because their frequency content is usually constrained to some frequency bands and do not affect all spectrum. Hence, if the above formulation is similarly applied to the artifacts $\mathbf{r}(k)$ then the combined signal-artifacts state-space model has a state vector

$$\mathbf{x}(k) = \begin{pmatrix} \mathbf{S}(k) \\ \mathbf{R}(k) \end{pmatrix} \tag{4}$$

and the transition matrix is

$$\mathbf{F}(k) = \left(\begin{array}{c|c} \mathbf{F}_s(k) & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{F}_r(k) \end{array}\right) \tag{5}$$

The remaining matrices are arranged in a comparable way.

In the described model the AR coefficients in the process matrix must be updated using the previous estimated values that result from the recursion. If the speech signal is processed by analysis of a set of overlapped windows then the spectral correlation between consecutive window is almost neglected. Also, depending on the window length and overlap ratio the time-frequency resolution may be insufficient for enhancing signal in the zones with higher frequency variations.

For improving time-frequency resolution, this work proposes a role inversion between the speech signal part of state-vector and the related part in the transition matrix. The new state-space vector is

$$\mathbf{x}(k) = \begin{pmatrix} \mathbf{a}(k) \\ \mathbf{r}(k) \end{pmatrix} \tag{6}$$

and the corresponding model is

$$\mathbf{x}(k) = \left(\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{F}_r(k) \end{array}\right) \mathbf{x}(k) + \mathbf{G} \begin{pmatrix} 0 \\ w(k) \end{pmatrix} \tag{7}$$

$$y(k) = \mathbf{H}\mathbf{x}(k) + v(k) \tag{8}$$

with

$$\mathbf{G}^T = \left(\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & 1 \underbrace{0 \ldots 0}_{N-1} \end{array}\right) \tag{9}$$

and

$$\mathbf{H}(k) = \left(\begin{array}{ccc|c} s(k) & \ldots & s(k-M-1) & 1 \underbrace{0 \ldots 0}_{N-1} \end{array}\right) \tag{10}$$

2.1. Dual Discrete Kalman filter

The Kalman filter, a particular case of the Wiener filter, can recursively estimate the state of a linear stochastic process such that the mean squared error is minimized. With the given model the AR parameters can be estimated in parallel with the state vector by two discrete Kalman filters.

The best linear estimate $\hat{\mathbf{x}}(k|k-1)$ at instant k using the knowledge up to instant $k-1$ is calculated as

$$\hat{\mathbf{x}}(k|k-1) = \mathbf{F}(k-1)\mathbf{x}(k-1|k-1) + \mathbf{G}\mathbf{w}(k-1) \tag{11}$$

and the related prediction error covariance matrix is

$$\mathbf{P}(k|k-1) = \mathbf{F}(k-1)\mathbf{P}(k-1|k-1)\mathbf{F}(k-1)^T + \mathbf{G}\mathbf{D}\mathbf{G}^T \tag{12}$$

where

$$\mathbf{D} = \begin{pmatrix} \sigma_s^2 & 0 \\ 0 & \sigma_r^2 \end{pmatrix} \tag{13}$$

and is time dependent. These values can be estimated directly from data since all the information is avalilable. Is some cases it is possible to force specific value since these variables act as confidence indicators and can used to make a more rigid or forgiving tracking.

The minimum mean square error (MMSE) of the state vector prediction is

$$\mathbf{S}(k) = \mathbf{H}(k)\mathbf{P}(k|k-1)\mathbf{H}^T(k) \tag{14}$$

which in this case is a value. The optimum Kalman gain is obtained from

$$\mathbf{K}(k) = \mathbf{P}(k|k-1)\mathbf{H}^T(k)\mathbf{S}(k)^{-1} \tag{15}$$

The new state estimate is now calculated considering the observed value and the prediction as

$$y(k) = x(n) \tag{16}$$

$$\hat{y}(k) = \mathbf{H}(k)\hat{\mathbf{x}}(k|k-1) \tag{17}$$

$$\hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1) + \mathbf{K}(k)\left[y(k) - \hat{y}(k)\right] \tag{18}$$

and the new covariance follows like

$$\mathbf{P}(k|k) = [\mathbf{I} - \mathbf{K}(k)\mathbf{H}(k)]\mathbf{P}(k|k-1) \tag{19}$$

where **I** is an identity matrix with size $M+N$.

The new AR coefficients estimate is now used to update the matrices $\mathbf{F}(k)$ and $\mathbf{H}(k)$. With this dual DKF algorithm it is possible to obtain instantaneous values for the speech signal and artifact estimation.

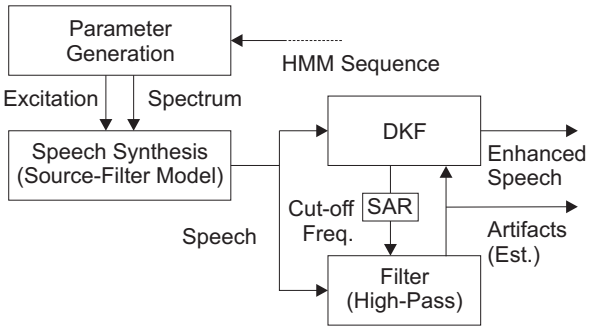

Fig. 1. Integration of the proposed DKF methodology in a typical HMM TTS architecture

2.2. PSD Modelling

The PSD estimates can be obtained directly from the state vector by

$$\hat{P}_x(e^{jw},k)=\frac{\left|\hat{b}(0,k)\right|^2}{\left|1+\sum_{i=1}^{M}\hat{a}_i(k)e^{-jwi}\right|^2} \qquad (20)$$

If consecutive instants are very near in time it is possible to average the model coefficients across several consecutive time frames for achieving more accurate results.

3. IMPLEMENTATION AND RESULTS

The used system architecture is presented on figure 1. The TTS output is decomposed, on a frame by frame basis, into excitation and filter using a 12th order AR model. The model estimation is based on the modified covariance method because it gives stable filter realizations and for high order models the extra poles are placed in a way that their effect is cancelled or convergent to the desired frequencies. The realizations are very accurate and the filter shows a high order robustness. The same TTS output is filtered using a variable cut-off frequency high pass filter which will provide the artefact signal estimation to which a similar AR modelling is applied. The PSD is estimated using both models and a signal to artifact ratio (SAR) is calculated for cut-off frequency adjustment. The best SAR value for our system was 20. It is initially assumed that the cut-off frequency of the filter is located at 5KHz.

The used HMM TTS has been built around the HTS [10] basic configuration (as provided in the distribution) using an European Portuguese voice font (30min hand labelled were used from a total of 60min). The audio was recorded in a professional studio with a professional male speaker using 16 bit samples and a 44100 hertz sampling rate. The utterances were extracted from newspaper and classic literature. For our purposes each sentence was saved in an independent file and the audio was downsampled to 16000 hertz . All the processing

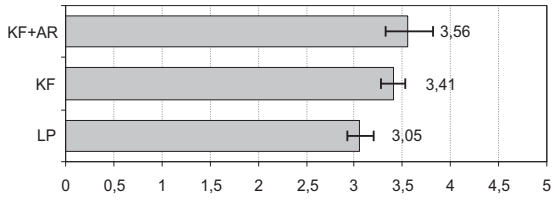

Fig. 2. Subjective performance evaluation on the MOS scale comparing

operations were performed with (Mathworks) Matlab scripts running under (Microsoft) Windows XP environment.

On a first experiment we tried to objectively evaluate the performance of the proposed methodology. A set of 10 sentences was randomly selected from the remaining 30min of the voice database. The same 10 sentences were generated using the TTS with a final 8KHz first order low-pass filter and their spectrum was compared with the original sentences using a discrete Itakura-Saito based metric [11]:

$$d_{IS}\left\{P(\omega),\hat{P}(\omega)\right\} = \frac{1}{N}\sum_{m=1}^{N}\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}-\log\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right) \qquad (21)$$

The obtained values served as a base line for the further comparisons. The same metric was used for the utterances generated using the Kalman filter only and using the Kalman filter with artifact estimation and the results showed 5% and 8% similarity improvement respectively.

Additionally a perceptive test has been made with 6 volunteers within the age range 19-23. The listeners were asked to classify 15 sentences, 5 low pass filtered (LP), 5 Kalman filtered (KF) and 5 Kalman filtered with artifact removal (KF+AR) according to intelligibility. A 1 to 5 points scale were used (1 for the worst result and 5 for the best result). As can be seen on figure 2 the KF+AR procedure achieved the best result very close to the KF only procedure. The 90% confidence interval of these last results is higher than then LP procedure. In both KF+AR and KF cases a minimum of 11.8% intelligibility increase was achieved when compared to the baseline LP system.

In the frequency domain we show an example in figure 3. The bottom spectrogram shows a more smoothed signal in time and the main formant frequency are still clearly defined. The overall energy of the signal is a bit reduced.

4. DISCUSSION AND FUTURE WORK

The obtained results are very interesting but the proposed approach is highly dependent on the used voice-font and the

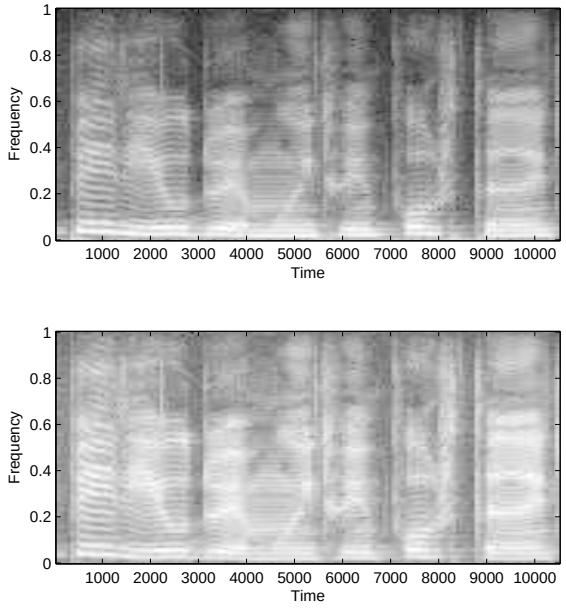

Fig. 3. Spectrograms for the original signal low pass filtered (top) and Kalman filtered with artifact removal (bottom). Both signals are the acoustical representation of the utterance "Período muito importante". The time (horizontal) axis is in sample units (using a 16KHz sampling frequency) and the frequency (vertical) axis is in normalized frequency (8KHz considered)

HMM training process and parameters. Some further experiments with females voices are still foreseen in order to tune the system.

The natural DKF tracking capabilities make it suitable to be used with different speech features. The line spectral frequencies (LSF) appear at specific angles in unitary circle that can be easily tracked. Using an independent Kalman filter for each LSF parameter can provide a better control of each frequency band. This work is on progress. The complete substitution of the speech generation filter by a Kalman based filter that directly receives the decoded speech parameters is also being worked.

5. CONCLUSION

In this work an adaptive modelling scheme based on Kalman Filtering is proposed for speech smoothing and artifact estimation. This filtering scheme is specifically developed for HMM based speech synthesis. It was shown how the speech and artifact estimation can be simultaneously integrated in a single state-space model. A dual Kalman filter algorithm is applied to this model in order to obtain improved signal estimates. The system's modelling ability and performance was successfully evaluated by objective and subjective tests.

6. REFERENCES

[1] R. Maia, H. Zen, K. Tokuda, T. Kitamura, and F. G. Resende, "Towards the development of a brazilian portuguese text-to-speech system based on hmm," in *Proc. of the European Conf. on Speech Communication and Technology (EUROSPEECH)*, 2003, pp. 2465–2468.

[2] Yamagishi J., Tachibana M., Masuko T., and Kobayashi T., "Speaking style adaptation using context clustering decision tree for hmm-based speech synthesis.," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[3] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. pp.2801–2804, 2005.

[4] R. Maia, Toda T., H. Zen, Nankaku Y., and K. Tokuda, "An excitation model for hmm-based speech synthesis based on residual modeling," in *Proc. of the ISCA Workshop on Speech Synthesis (SSW'06)*, Bonn, 2007, pp. 131–136.

[5] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall, Englewood Cliffs, NJ, 2000.

[6] E. Zavarehei and S. Vaseghi, "Speech enhancement in temporal dft trajectories using kalman filters," in *Proc. of Interspeech 2005*, Lisboa, 2005, pp. 2077–2080.

[7] S. Boll, "Supression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[8] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. of the Seventh European Signal Processing Conference (EUSIPCO-94)*, Edinburgh, 2007, pp. 1182–1185.

[9] E. A. Wan and R. Merwe, "Noise-regulated adaptive filtering for speech enhancement," in *Proc. of Eurospeech 99*, Budapest, 1999.

[10] HTS, "Hmm-based speech synthesis system (hts)," June 2007, at http://hts.sp.nitech.ac.jp/.

[11] R. J. McAulay, "Maximum likelihood spectral estimation and its application to narrow-band speech coding," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 32, no. 2, pp. 243–251, 1984.