$$\mu_i = \frac{\sum_m w_m^i E(x_m | Q_m = i,\ D_m)}{\sum_m w_m^i} \qquad (11)$$

$$\sum_i = \frac{\sum_m w_m^i E(x_m x_m' | Q_m = i,\ D_m)}{\sum_m w_m^i} - \mu_i \mu_i' \qquad (12)$$

where function $E(x)$ is the expected likelihood of x, $w_m^i = E(q_m^i | D_m)$, the variable $q_m^i = 1$ if Q has value i in the mth data cases, and 0 otherwise. A more general learning method over various conditional probability distributions has been described in [3].

Testing: The aim in the testing procedure of speaker identification is to determine the right person given an observation, $\hat{i} = \arg_i \max p(M_i|O)$, $i = 1, \ldots, N$, where M_i is the model of speaker i. It means we have to calculate the posterior probability, $p(M_i|O)$, $i = 1, \ldots, N$. According to the Bayes rule, $p(M|O) = p(O|M) * p(M)/p(O)$. Since no knowledge about the prior probability $p(O)$ is known and probability $p(M)$ is the same for all the models, we use $p(O|M)$ in substitute of $p(M|O)$ for simplicity. It can be achieved by computing the joint probability using (9).

Experiments: In our experiments, we defined the topology of the DBNs as shown in Fig. 1, which is unrolled for first two slices. q_j^i, $i = 1, 2, 3, j = 1, 2, \ldots, T$ are hidden nodes and have discrete values, o_j^i, $i = 1, 2, 3, j = 1, 2, \ldots, T$ can be observed and satisfy Gaussian distributions, here T is the length of time slices.

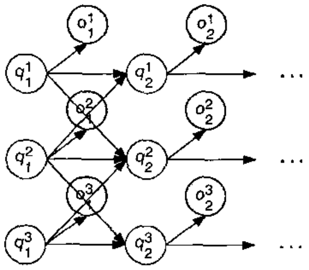

Fig. 1 *Topology of DBNs used in experiments*

For computational reasons, a subset of the YOHO database including the first 30 speakers (speaker ID from 101 to 132, excluding 123 and 129 since no data of these two speakers exist in YOHO corpus) is used, and only the first session (24 sentences) is used for training. In the feature selection, the Hamming window size is 32 ms (256 samples), and silence and unvoiced segments are discarded based on an energy threshold. The feature vector is composed by 16 MFCC. All the verify sentences (40 per speaker) are used in the test procedure. We compare the recognition rates over different numbers of speakers of DBNs to other classical methods such as Gaussian mixture model (GMM), continuous hidden Markov model (CHMM), single Gaussian and vector quantisation (VQ) in Table 1. It can be seen that the DBNs-based method has achieved encouraging results.

Table 1: Recognition rate results under various methods over different number of speakers

Method	Number of speakers		
	First 10 (%)	First 20 (%)	First 30 (%)
DBNs	100	97.2500	96.6667
GMM	99.5000	95.1250	94.7500
CHMM	99.2500	94.1250	94.4167
Single Gaussian	99.0000	94.0000	92.6667
VQ	98.5000	94.6250	94.0833

Some parameters of models in experiments: GMM-16 mixtures; CHMM-5 states and 10 GMM outputs; VQ-128 codebook size

Conclusion: Dynamic Bayesian networks are expressive models for stochastic processing. We have presented a framework using dynamic Bayesian networks for speaker identification and have achieved considerable results in the experiments on a subset of YOHO corpus.

Acknowledgments: We wish to acknowledge support for this work by the National High Technology Research & Development Programme (863) of the People's Republic of China (No. 2001AA4180) and the Zhejiang Provincial Natural Science Foundation for Young Scientists of the People's Republic of China (No. RC01058).

© IEE 2003 *26 October 2002*
Electronics Letters Online No: 20030184
DOI: 10.1049/el:20030184

Lifeng Sang, Yingchun Yang, Zhaohui Wu and Wanfeng Zhang (*Department of Computer Science & Engineering, Zhejiang University, Hangzhou, 310027, People's Republic of China*)

E-mail: lfsang@cs.zju.edu.cn

References

1 ZWEIG, G.G.: 'Speech recognition with dynamic Bayesian networks'. PhD Thesis, University of California, Berkeley, CA, USA, 1998

2 COWELL, R.: 'Introduction to inference for Bayesian networks' *in* JORDAN, M.I. (Ed.): 'Learning in graphical models', (MIT Press, 1999), pp. 9–26

3 MURPHY, K.: 'Dynamic Bayesian networks: representation, inference and learning'. PhD Thesis, University of California, Berkeley, CA, USA, 2002

Real-time room acoustic response simulation by IIR adaptive filter

G. Costantini and A. Uncini

A new IIR adaptive filter for real-time, room acoustic response simulation is proposed, the structure of which derives from Jot's model of an artificial reverberator. The simultaneous perturbation stochastic approximation (SPSA) algorithm is used to set parameter values. Results show good similarity between the desired and artificial response.

Introduction: It is common knowledge that different types of music require different acoustical characteristics, related to their particular requirements. Every closed space induces some reflections of the signal generated inside and, together with other phenomena, have influence on the auditory feeling of a listener. For example, listening to a sacred music choir requires a very reverberant place, that stirs the complex of vocal sounds, and works like a powerful case of resonance; a piano concert, however, requires a drier ambient, that allows careful distinction of every single note. Therefore, there exists the problem of rendering a listening place adaptable to different types of music, artificially reproducing acoustical characteristics of a particular ambient. Our research objective is to obtain an accurate reproduction of an acoustical impulse response (IR) of a generic room through a real-time musical signal processing, introducing similar reverberation and spectral characteristics of a target ambient.

Reverberation: The reverberation phenomenon consists in a persistence of sound gradually attenuated during a certain temporal interval after the sound source has stopped. This is due to the multiple wall reflection of acoustic spherical waves in the listening room, which reflects them with an absorption coefficient α, dependent on the constitutive materials [1]. Reflected waves can be divided into two classes. First, there are the early reflections, i.e. waves reflected only once before reaching the listener; this part is very important to create in the listener's brain the idea of the room spatial dimensions. Then, many other wave-rays are reflected more times by the walls before reaching the listener, thus creating a very dense whole of echoes that arrive at the listener in random times and constitute the reverberating tail.

To quantify the reverberation entity in a room, the reverberation time T_{60} is defined as the time for the impulse response decaying from -5 to

−65 dB of its maximum level. An ambient is very reverberant if $T_{60} > 2$s, while it is very dry if $T_{60} < 1$s.

Artificial reverberation: Schroeder was the pioneer who first attempted to make digital reverberation. The first prototype he tried, called comb filter (or *plain reverberator*), consisted of a single delay line of m samples with a feedback loop containing an attenuation gain g. To simulate the frequency selective absorption of air and walls, Schroeder modified the plain reverberator, inserting a lowpass filter in the feedback loop; in order to increase the time echoes density and to decrease the metallic effect produced by comb filters, he cascaded multiple all-pass filters. Schroeder's structure is composed of four parallel comb filters, followed by two cascaded all-pass ones [2]. The Schroeder frequency response results provide a good quality, but it is rather 'anonymous'.

An evolution of Schroeder's model was developed by Moorer [3]. He observed that Schroeder's model can well modelise only the reverberation tail, while a room is characterised by its early reflections, that have to be reproduced by an FIR filter. Thus Moorer added an FIR filter before the Schroeder reverberator with the aim of efficaciously reproducing the early reflections. Moorer chose the length of this FIR filter to reproduce the first 60–80 ms of the real impulse response.

Employed structure: The Moorer and Schroeder models can be used to provide a well-sounding effect, setting by hand the coefficient of the basic blocks (comb and all-pass) that compose them. Our objective is quite different: we seek to obtain a general purpose effect; we want to simulate artificially, as faithfully as possible, the acoustic response of a particular room. Starting by a defined filtering structure and a desired IR, we have to find a procedure to correctly identify the coefficients of the filter, so that it can satisfy the following:

– the reverberated signal sounds well, i.e. it comes out pleasurable for the subjective hearing;
– the dry signal filtered by the identified reverberator is as similar as possible to the signal really produced inside the room, the IR of which is measured;
– the structure is purposed for any real-time implementation.

The structure chosen for the artificial reverberator is shown in Fig. 1. It is a generalisation of the filtering structure developed by Jot in 1992 [4].

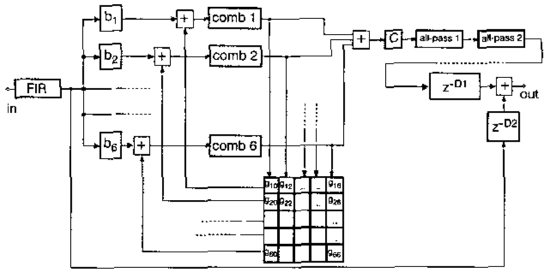

Fig. 1 *Generalised Jot reverberator flow diagram*

A structure that can efficaciously simulate a desired IR must have the properties of:

Elasticity: the filter must have quite a number of free parameters and must allow free variations of its characteristics;
Robustness: the filter must be stable and able to produce every kind of IR, with different T_{60};
Enough complexity: the structure must generate an high echo density.

A structure these satisfy these properties is an 'adaptive filter';

The novelty that Jot introduced, compared to the Moorer model, is the presence, in the IIR part, of a feedback coefficients matrix **G**, this being for two main reasons:

– remixing in time the echoes, the aim being to cancel the periodicity effect introduced by the comb filters and also by the FIR filter;
– increasing the number of free parameters for a better approximation of the real IR.

For the **G** matrix we chose $g_{ii} = 0$ ($i = 1, \ldots, 6$), so every comb is fedback on all the others, except on itself. The b_i and c coefficients are employed for scaling every comb and all-pass contribution, thus increasing the elasticity of the filter. Thus the employed structure is characterised by the following parameters:

– g_{i1}, g_{i2}, D_i for the six comb filters (18 coefficients);
– the **G** matrix (36 coefficients);
– b_i and c (7 coefficients);
– g_i and R_i for the two all-pass filters (4 coefficients);

for a total number of 65 free coefficients. We can indicate the transfer function of the whole structure with $F(z, \hat{\mathbf{w}})$, where $\hat{\mathbf{w}}$ indicates the free coefficients vector.

Identification procedure: The identification procedure requires finding the values of the 65 coefficients that make the IR of the filter as similar as possible to the desired IR. It is impossible to accomplish this procedure by setting the parameters by hand; we have to employ, on the contrary, an automatic procedure, in which the coefficients of the filter are iteratively corrected by a suitable adaptation algorithm, based on the difference $e(n)$ between the desired response $d(n)$ and the filter output response $\hat{d}(n)$, to make the error $e(n)$ as low as possible.

The optimisation algorithm used for the reverberator filter identification is the simultaneous perturbation stochastic approximation (SPSA); it was developed by Spall in 1992 [5]. Let us consider the problem of minimising a scalar differentiable loss function $J(\hat{\mathbf{w}})$, where $\hat{\mathbf{w}}$ is the p-dimensional vector of parameters: the optimisation problem can be translated into finding the minimising $\hat{\mathbf{w}}^*$ such that $\partial J/\partial\hat{\mathbf{w}} = 0$.

The choice of the loss function $J(\cdot)$ represents the crucial point of the entire identification procedure: the final result depends on the capacity of the loss function to express the filter reverberation quality, also regarding the sentence 'it well sounds'. $J(\cdot)$ must quantify the similarity between the artificial response and the real response, as well as guaranteeing a good quality of artificial reverberation.

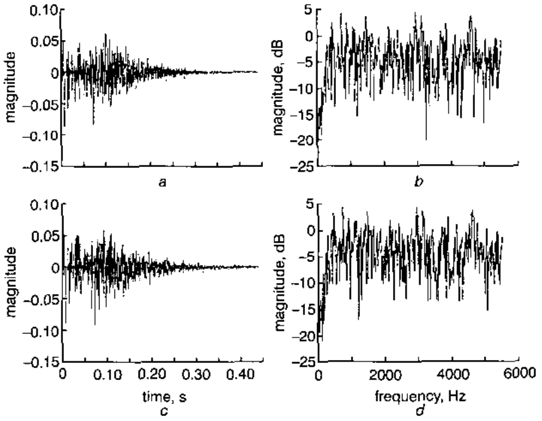

Fig. 2 *Comparison between desired and artificial response*

a Desired time response
b Desired frequency response
c Artificial Time response
d Artificial frequency response

The loss function we propose in this Letter carries out:

(i) A comparison between the power $p(t)$ of simulated IR and real IR, to re-create the envelope and the echo distribution; we windowed the temporal axis in N windows, calculated the power in each window and minimised the maximum power. It allows us to simulate rather well the temporal behaviour of the real IR, reproducing it with the same T_{60} and echo distribution.
(ii) A comparison between the real and the artificial frequency responses, minimising the mean square error, to obtain a similar frequency colouration between the real and artificial entire frequency response.
(iii) A further frequency test, minimising the maximum shifting, to reduce the defects introduced by the employed structure. We windowed

the frequency axis in M windows; then minimised the maximum error between the frequency responses in each window.

Simulation results: The adaptive filter described above has been tested by reproducing impulse responses artificially generated by an acoustical editing tool based on ray tracing and image-source methods. We set this tool to generate typical IR, assuming rectangular perpendicular rooms.

We report the results related to a small ambient, with a medium wall absorption coefficient $\alpha = 0.95$. Figs. 2*a* and *b* show the time response and the frequency response, respectively, of this room, assuming a sample frequency $f_s = 11025$ Hz. For the FIR filter, we chose the first IR 60 ms. At the end of the identification procedure, we obtained the filter responses shown in Figs. 2*c* and *d*. A good similarity between the desired and the artificial time and frequency response can be observed: the artificial reverb reached the same T_{60} and the same harmonic content of the desired test room responses.

© IEE 2003 *9 December 2002*
Electronics Letters Online No: 20030185
DOI: 10.1049/el:20030185

G. Costantini (*University of Rome 'Tor Vergata' — Department of Electronics Engineering, Via del Politecnico, 1-00133 Roma, Italy*)

E-mail: costantini@uniroma2.it

A. Uncini (*University of Rome 'La Sapienza' — INFO-COM Department, Via Eudossiana, 18-00184 Roma, Italy*)

References

1 SABINE, W.C.: 'Reverberation' *in* LINDSAY, R.B. (Ed.): 'Acoustics: historical and philosophical development' (Dowden, Hutchinson, and Ross, Stroudsburg, PA, USA, 1972, first published 1900)
2 SCHROEDER, M.R.: 'Natural sounding artificial reverberation', *J. Audio Eng. Soc.*, 1962, **10**, (3), p. 219
3 MOORER, J.A.: 'About this reverberation business', *Comput. Music J.*, 1979, **3**, (2), pp. 13–28
4 JOT, J.M.: 'An analysis/synthesis approach to real-time artificial reverberation'. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, San Francisco, CA, USA, 1992, Vol. 2, pp. 221–224
5 SPALL, J.C.: 'Multivariate stochastic approximation using a simultaneous perturbation gradient approximation', *IEEE Trans. Autom. Control*, 1992, **37**, pp. 322–341

Simplification of soft-bit speech decoding and application to MELP encoded speech

Xiaobei Liu and Soo Ngee Koh

A simplified soft-bit speech decoding (SBSD) algorithm is proposed and is used in decoding the encoded speech parameters from the mixed excitation linear prediction (MELP) coder. Simulation results show that the simplified algorithm can reduce significantly the decoding complexity and memory requirement of SBSD at a slight penalty of a marginal drop in performance.

Introduction: Soft-bit speech decoding (SBSD), as a method which successfully uses the residual redundancy to achieve error concealment, can be applied to almost all speech coding algorithms with significant improvements in the quality of the decoded speech when the encoded bits are received with errors [1–3]. The main reason that deters the use of SBSD in real-time applications is the exorbitant complexity and memory requirement of the decoding algorithm, especially for parameters that are encoded by a large number of bits.

The complexity issue is only considered in [2] in which the authors propose to employ lower dimensional vector quantisers (VQs). However, this approach requires the redesign of codebooks at the encoder end. In this Letter, a simple but effective simplified SBSD without modifying the encoder is proposed and is applied to mixed excitation linear prediction (MELP) encoded speech parameters [4, 5]. Simulation results show that while our proposed method reduces significantly the decoding complexity and memory requirement, it only costs a slight degradation in the quality of the decoded speech.

Complexity of SBSD: The SBSD algorithm can be divided into two steps. First, the decoder uses the received sequence $\underline{\hat{X}}_k = \{\underline{\hat{x}}_1, \underline{\hat{x}}_2, \ldots, \underline{\hat{x}}_k\}$ to compute the *a posteriori* probability (APP) of each of the possible transmitted bit combination $\underline{x}_k^i = \{x_k^i(0), x_k^i(1), \ldots, x_k^i(M-1)\}$, where i is the index of the parameter v_k and $i \in \{0, 1, \ldots, 2^M - 1\}$. The resulting equation is:

$$P(\underline{x}_k^i|\underline{\hat{X}}_k) = C \cdot P(\underline{\hat{x}}_k|\underline{x}_k^i) \sum_{j=0}^{2^M-1} P(\underline{x}_k^i|\underline{x}_{k-1}^j) \cdot P(\underline{x}_{k-1}^j|\underline{\hat{X}}_{k-1}) \qquad (1)$$

where C is the normalisation constant. Secondly, the APPs are used for MMSE estimation of the reconstructed parameter $\hat{v}_k$ as shown in (2):

$$\hat{v}_k = \sum_{i=0}^{2^M-1} v_k^i \cdot P(\underline{x}_k^i|\underline{\hat{X}}_k) \qquad (2)$$

From the above, we can see that the total complexity for calculating each parameter is about 2^{2M} for both multiplication and addition. The memory required is 2^{2M} also for storing the *a priori* index transition probabilities $P(\underline{x}_k^i|\underline{x}_{k-1}^j)$.

Simplified SBSD: Suppose $\underline{x}_k^{i'}$ is the combination of the first N significant bits of $\underline{x}_k^i$, which means $\underline{x}_k^{i'} = \{x_k^i(M - N), x_k^i(M-N+1), \ldots, x_k^i(M-1)\}$, $N \leq M$. To reduce the memory requirement of SBSD, we propose to divide the 2^M parameter indices into 2^N groups and each group includes 2^{M-N} of $\underline{x}_k^i$ with the same $\underline{x}_k^{i'}$. The transition probabilities between groups are used instead of the individual index transition probabilities. The transition probability between each two groups is calculated as

$$P(\underline{x}_k^{i'}|\underline{x}_{k-1}^{j'}) = \frac{\sum_{i,\underline{x}_k^{i'}=i'} \sum_{j,\underline{x}_{k-1}^{j'}=j'} P(\underline{x}_k = i, \underline{x}_{k-1} = j)}{\sum_{j,\underline{x}_{k-1}^{j'}=j'} P(\underline{x}_{k-1} = j)} \qquad (3)$$

Equation (1) now becomes

$$P(\underline{x}_k^i|\underline{\hat{X}}_k) = C \cdot P(\underline{\hat{x}}_k|\underline{x}_k^i) \sum_{j'=0}^{2^N-1} P(\underline{x}_k^{i'}|\underline{x}_{k-1}^{j'}) \cdot P(\underline{x}_{k-1}^{j'}|\underline{\hat{X}}_{k-1}) \qquad (4)$$

where

$$P(\underline{x}_{k-1}^{j'}|\underline{\hat{X}}_{k-1}) = \sum_{j,\underline{x}_{k-1}^{j'}=j'} P(\underline{x}_{k-1}^j|\hat{X}_{k-1}) \qquad (5)$$

From the above, it can be found that the memory requirement is reduced to 2^{2N} and the numbers of multiplications and additions are 2^{M+N} and 2^{2M}, respectively. To further reduce the number of calculations, we propose to ignore the 'index transmission probabilities' $P(\underline{\hat{x}}_k|\underline{x}_k^i)$ [1] below a certain value T since they contribute negligibly to the reconstructed parameters. Suppose R out of the 2^M index transmission probabilities remain after ignoring the others, the numbers of multiplications and additions then become less than $R \cdot 2^N$ and $R \cdot 2^M$, respectively.

Results and discussion: To test the performance of the proposed procedure, the simplified SBSD is first applied to the MELP encoded pitch parameter [6] which is quantised to 7 bits and requires 16 384 arithmetic operations and words. Table 1 shows the parameter SNR (PSNR) of pitch at $E_b/N_o = 0$ dB obtained by the simplified SBSD for different values of T and R. It should be noted that the PSNR of pitch with the unmodified SBSD is 16.6 dB and its value without the use of SBSD is only 9.77 dB.

From Table 1, it can be observed that the performance improves with increasing value of N and decreasing value of T. However, when $N > 4$ or $T < 10^{-3}$, the amount of improvement becomes marginal whereas the number of arithmetic operations increases rapidly. When $T = 10^{-5}$, there is even a slight decrease in the PSNR, which can be explained by the indices with the transmission probabilities below 10^{-4} being very unlikely to be the correct transmitted indices and taking them into account in the reconstruction of the parameters will only degrade the decoder performance.