FREQUENCY RECOVERY OF NARROW-BAND SPEECH USING ADAPTIVE SPLINE NEURAL NETWORKS

Aurelio Uncini, Francesco Gobbi and Francesco Piazza

Dipartimento di Elettronica e Automatica - Università di Ancona Italy
Via Brecce Bianche, 60131 Ancona-Italy
Phone :+39 (71) 2204841- Fax:+39 (71) 2204464
email: aurel@eealab.unian.it - URL: http://nnsp.eealab.unian.it/

ABSTRACT

In this paper a new system for speech quality enhancement (SQE) is presented. A SQE system attempts to recover the high and low frequencies from a narrow-band speech signal, usually working as a post-processor at the receiver side of a transmission system. The new system operates directly in the frequency domain using complex-valued neural networks. In order to reduce the computational burden and improve the generalization capabilities, a new architecture based on a recently introduced neural network, called adaptive spline neural network (ASNN), is employed. Experimental results demonstrate the effectiveness of the proposed method.

1. INTRODUCTION

In the last years the speech quality enhancement (SQE) systems have drawn a growing interest since they can improve the unnatural feel of narrow-band telephone speech. In the past already the British Broadcasting Corporation (BBC) performed several experiments of narrow-band speech enhancement, using noise generators and non linear speech processing [1]. More recently several authors have proposed methods based on linear and non linear digital processing of time-frequency speech signals [2-7].

In [2], the authors propose a spectral envelope extrapolation based on a linear predictive coding (LPC) approach. A codebook consisting of wide-band LPC envelopes must be available. During the enhancement process, the short-time spectra of the narrow-band speech signal to be enhanced are computed every 20 ms. For each frame, the best-fitting entry of the codebook is selected as the desired wide-band envelope estimate. The fine spectral structure is built using a simple excitation signal derived by processing the narrow-band signal with the so-called High-Frequency Regenerator (HFR) which performs a simple spectral folding.

In [3], Hermansky et al. propose the use of a bank of Wiener-like non-causal FIR filters for the prediction of cubic-root compressed short-time power spectrum. This non-linear technique is based on the RelAtive SpecTrAl (RASTA) processing of speech firstly used to enhance the quality of noisy cellular telephone communications. The main drawback of this approach is in the phase reconstruction which is performed using the low frequency signal, producing annoying musical-like residual noise.

In [4] the same authors proposed a different predictive approach. The spectral predictor is realized with a linear all-pole filter. The envelope prediction is performed by filtering the time trajectory of LPC-cepstral coefficients of the narrow-band speech signal with a multidimensional filter designed on some training data. The fine spectral structure is derived by an excitation signal obtained by a technique used in codebook-excited speech coding.

A very simple approach, suitable for real time application on a low-cost DSP, is proposed in [5]. The narrow-band speech signal is enhanced by simple non linear processing. The non linear process consists in the cascade of a rectifier, an high pass filter, and a shaping filter followed by a block which performs a level adjustment tuned on a subjective assessment.

In [6] this system is improved using an adaptive filter which performs the high frequency spectral shaping and the level adjustment. Other improvements can by found in where the LPC parameters of the narrow-band signal are used to produce the wide-band parameters using a simple layered neural network.

In this paper we proposed a simple neural scheme to perform the wide-band frequency recovery from a narrow-band speech. Neural network approach, in fact, allows to extract the missing frequency contents by a simple non-linear mapping, in the frequency domain, from narrow to broad band speech signal.

The proposed SQE system works without an excitation signal and does not need any parameter tuning.

The computational load of standard neural networks is overcome using a recently proposed architecture, very suitable for signal processing application, based on an adaptive spline activation function called adaptive spline neural network (ASSN) [8-10].

2. THE NEURAL NETWORK SPEECH QUALITY ENHANCEMENT SYSTEM

2.1 Problem definition

It is known that a signal sampled at 16KHz (wide-band speech) has a nominal frequency band from 0 to 8KHz, while the narrow band telephone speech is limited between 300 and 3400 Hz. The problem is therefore to recover from this narrow band signal the two missing frequency bands: nominally from 0 to 300 Hz and from 3400 to 8000 Hz. As supposed by other authors [2-7], this should be made possible by the human speech production mechanism, which relates the frequency contents of different bands.

Let $s[n]$ be the narrow-band speech signal whose short-time Fourier transform (STFT) is $S_n\left(e^{j\omega_k}\right)$, with $S_n\left(e^{j\omega_k}\right) \neq 0$ only for $\omega \in [\omega_1, \omega_2]$. Let $\tilde{s}[n]$ be the corresponding wide-band signal; its STFT is now $\widetilde{S}_n\left(e^{j\omega_k}\right)\neq 0$ for $\omega\in\left[\omega_0,\omega_N\right]$ with the position $\omega_0<\omega_1$ and $\omega_2<\omega_N$.

A SQE system assumes the existence of an operator Ψ (in general non-linear), called quality enhancement operator (QEO), such that:

$$\widetilde{S}_n\left(e^{j\omega_k}\right)=\Psi\left[S_n\left(e^{j\omega_k}\right)\right] \tag{1}$$

or, in terms of STFT:

$$\begin{aligned}\widetilde{s}[n]&=\sum_m\left[\sum_k\Psi\left[S_n\left(e^{j\omega_k}\right)\right]e^{j\omega_k n}\right]\\&=\sum_m\left[\sum_k\Psi\left[\sum_l s[l]w[m-l]e^{-j\omega_k l}\right]e^{j\omega_k n}\right]\end{aligned} \tag{2}$$

where w[.] represents the overlapping time window.

2.2 The SQE Architecture

Basically, the proposed SQE system performs a direct STFT on a narrow-band signal upsampled to 16KHz, recovers the broad-band signal through the non-linear operator Ψ, and performs an inverse STFT.

Such a simple scheme, however, does not attain good performances, since the recovery processes for the lower and the higher band are very different. Moreover the Ψ operator could degrade the narrow-band frequency contents of the original speech signal. Better performances hence can be obtained by splitting the Ψ operator in two different operators Ψ_L and Ψ_H, one for the lower frequencies and the other for the higher ones. The original narrow-band information are sent to the output without any processing. A detailed scheme of such a SQE system is shown in figure 1:

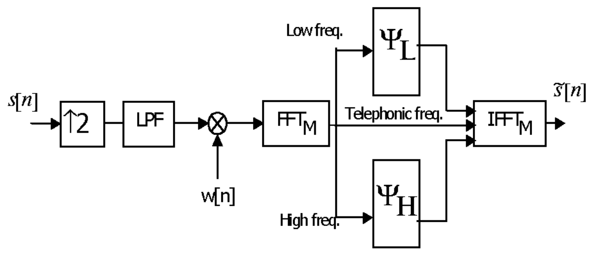

Figure 1. The architecture of the proposed enhancer (M is the length of the window w[n] used for the STSF and LPF is the interpolation filter).

The proposed system implements both the Ψ_L and Ψ_H operators with properly trained complex adaptive spline neural networks, respectively ASNN1 for the first operator and ASNN2 for the second.

However, since it is known that the frequency contents of the higher band (3600-8000 Hz) is strongly related to the contents of the narrow-band frequencies mainly for voiced sounds, our SQE system processes differently high frequency voiced and unvoiced sounds. For the first the ASNN2 properly trained only on voiced speech is employed, while for the latter the scheme [5] of figure 2 is also used.

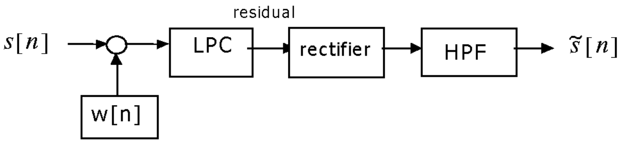

Figure 2. Additional scheme for recovery of high frequency unvoiced sounds (HPF is a high pass filter as in [5]).

2.3 Neural Network Architecture and Learning

In order to develop a real-time SQE system, a specific NN architecture with adaptive activation function is used [8-10].

This network is designed using a neuron containing an adaptive parametric spline activation function. The multilayer networks built with such neurons are still universal approximators and have usually a smaller structural complexity, maintaining good generalization capabilities.

The spline activation functions are smooth parametric curves, divided in multiple tracts. The i-th tract of the curve $\boldsymbol{F_i}(u)$ is represented by

$$F_i(u)=\left[F_{xi}(u),F_{yi}(u)\right]^T\ ;\quad 0\leq u\leq 1 \tag{3}$$

where u is the parameter, T is the transpose operator and $F_{xi}(.)$, $F_{yi}(.)$ are two polynomial functions describing the curve in the two coordinates x and y.

In particular due to the continuous first derivative, which allows to develop a backpropagation-style learning algorithm [8], we use the Catmull-Rom-based spline and the expression (3) is simply rewritten as

$$F_i(u)=\left[u^3\ u^2\ u^1\ 1\right]\frac{1}{2}\begin{bmatrix}-1&3&-3&1\\2&-5&4&-1\\-1&0&1&0\\0&2&0&0\end{bmatrix}\begin{bmatrix}Q_i\\Q_{i+1}\\Q_{i+2}\\Q_{i+3}\end{bmatrix}\ ; \tag{4}$$

where $u\in[0,1]$ and Q_{i+m} m=0,...,3 are the four *control points* for each curve tract. Such spline schemes are called *local schemes*, as the shape of the curve in the i-th tract is affected *only* by its four control points, so that the curve can be locally modified without influencing distant tracts. Eq. (4) represents the neuron output for the i-th tract.

The proposed neuron is composed of a classical linear combiner, which performs the weighted sum of the inputs, and of two blocks (SG1 and SG2) which implement the spline adaptive activation function (see Figure 3).

The block SG1 performs the mapping of the linear combiner output to the parametric domain, while the block SG2 computes the neuron output by using the activation function's control points, stored in a lock-up table (LUT), and the polynomial coefficients of Eq. (4). The learning algorithm and the extension to complex-valued signals can be found in [8-10].

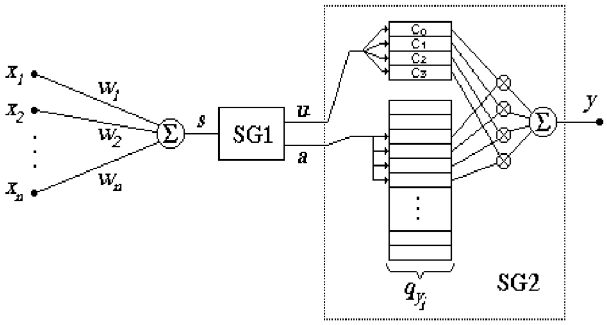

Figure 3. The neuron scheme based on the Catmull-Rom spline adaptive activation function with the internal structure of the SG2 block

In our architecture, both ASSN1 and ASNN2 are complex-valued and trained using a database of several speech signals sampled at 16KHz, following the scheme shown in Figure 4. The ASSN2 however is trained only on voiced frames through the use of a simple voiced/unvoiced selector.

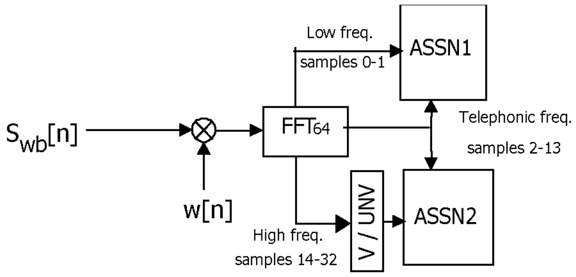

Figure 4. Scheme of the system used for training the two neural networks. V/UNV is a voiced/unvoiced selector and $s_{wb}[n]$ is the original wide-band speech signal.

3. EXPERIMENTAL RESULTS

Several different experiments have been carried out with various combinations of networks and signals. Here are briefly presented some results obtained with the following experimental setup:

- 64 point Hanning windows with an overlap of 32 points;
- 64 points complex FFT/IFFT;
- a complex ASNN1 with 12 inputs and 2 outputs, with $\Delta x=1$ and 40 control points for each neuron [8] for the recovery of the lower band;
- a complex ASNN2 with 12 inputs and 19 outputs, with $\Delta x=1$ and 40 control points for each neuron [8] for the recovery of the higher band (voiced sounds);
- an LPC of order 8 for the recovery of the higher band (unvoiced sounds) followed by a rectifier and a high-pass filter;
- several male and female telephone speech signals.

Some results in terms of SNR (Signal-To-Noise ratio considering as noise the difference between the original wide-band speech and the signal itself) are reported in Table 1. Figure 5 shows the time-frequency plots of the original narrow-band signal, the original wide-band signal and the output of the proposed system. Note that the low frequency recovery is very good, while the recovery of the high frequency contents is less efficient especially when unvoiced sounds are involved. This behavior can be better realized in Figure 6 that shows the typical spectra of particular voiced and unvoiced sounds.

The proposed system is able to sensibly improve the quality of the perceived speech, as indicated by the SNR indexes. The output speech sounds much better than the original telephone counterpart, although not as good as the original wide-band version.

TABLE 1. Performance of the proposed SQE system: global Signal-To-Noise ratio (SNR), segmented with 64 point windows (SNR segm), maximum (SNR max) and minimum SNR (SNR min) over all the windows.

	Original narrow-band speech	Output of the SQE system
SNR	2.688722 dB	14.695422 dB
SNR segm	2.715295 dB	12.903535 dB
SNR max	23.931820 dB	39.284490 dB
SNR min	-11.765533 dB	-10.252285 dB

4. REFERENCES

[1] M. G. Croll, «Sound Quality Improvement of Broadcast Telephone Calls», *BBC Research Report RD1972/26*, British Broadcasting Corporation, 1972.

[2] H. Carl and U. Heute «Bandwidth Enhancement of Narrow-Band Speech Signals», *Proc. of EUSIPCO 1994*, pp. 1178-1181, 1994.

[3] H. Hermansky, E. A. Wan, and C. Avendano, «Speech Enhancement based on Temporal Processing», *Proc. of IEEE ICASSP95,* Detroit, MI, USA, pp. 405-408, May 95.

[4] C. Avendano, H. Hermansky, and E. A. Wan, «Beyond Nyquist: Towards the Recovery of Broad-band Speech from narrow-bandwidth Speech», *Proc. of EUROSPEECH'95*, pp. 165-168, Sept. 1995.

[5] H. Yasukawa, «Signal Restoration of Broad Band Speech Using Nonlinear Processing», *Proc. of EUSIPCO'96*, Trieste, Italy, Sept. 1996.

[6] H. Yasukawa, «Adaptive Digital Filtering For Signal Reconstruction Using Spectrum Extrapolation», *Proc. of EUSIPCO'96*, Trieste, Italy, Sept. 1996.

[7] H. Yasukawa, «Wideband Speech Recovery from Bandlimited Speech in Telephone Communications», *Proc. of IEEE ISCAS'98*, Monterey, CA, USA, May. 1998.

[8] L. Vecci, F. Piazza, A. Uncini, "Learning and approximation Capabilities of Adaptive Spline Activation Function Neural Networks", *Neural Networks*, Vol. 11, No. 2, pp.259-270, March 1998.

[9] A. Uncini, F. Capparelli, F. Piazza, "Fast Complex Adaptive Spline Neural Networks for Digital Signal Processing", Proc. of IJCNN'98, Anchorage (Alaska), pp. 903-909, May 1998

[10] A. Uncini, L. Vecci, P. Campolucci, F. Piazza, "Complex-valued Neural Network with Adaptive Spline Activation Function for Digital Radio Links Nonlinear Equalization" to appear on *IEEE Trans. On Signal Processing*

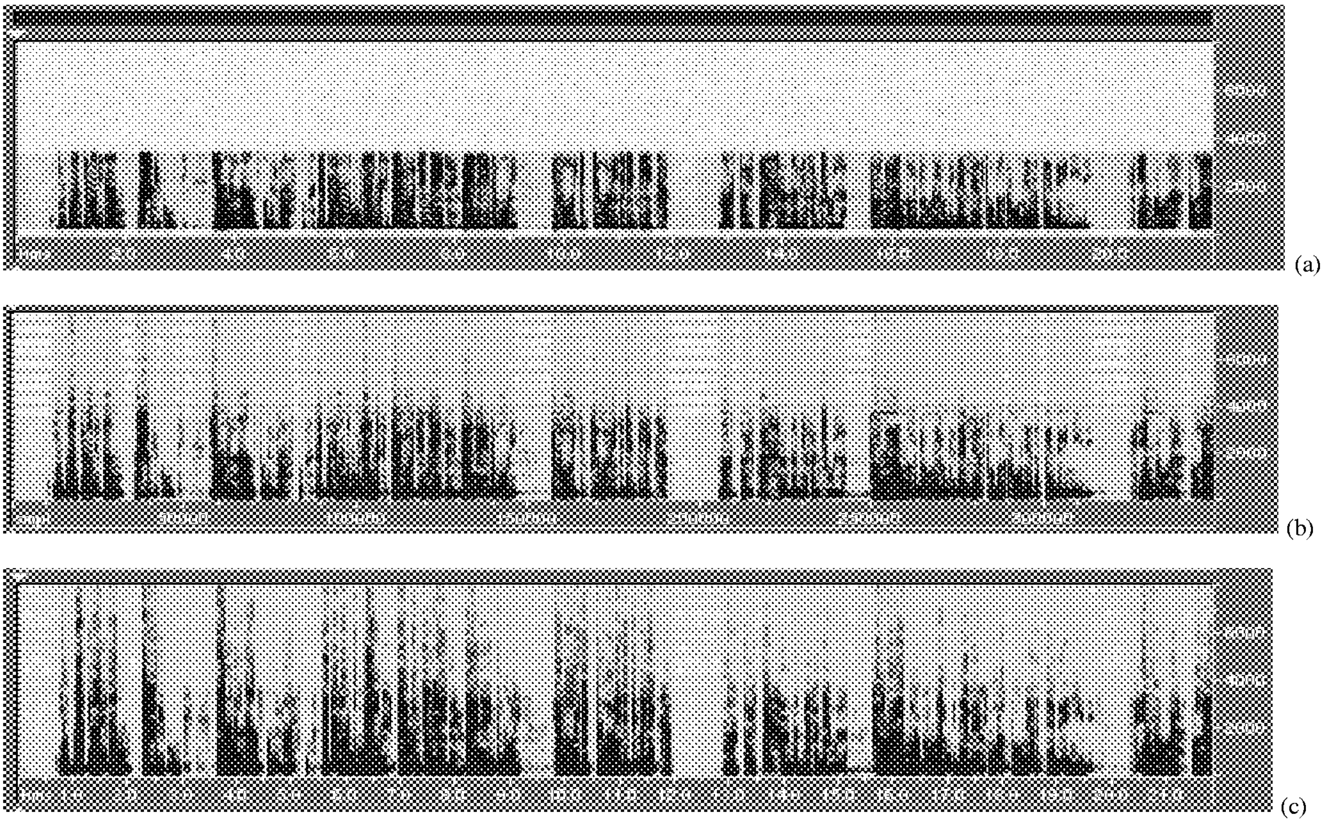

Figure 5. Time-frequency plots of speech signal containing utterances from different female speakers: (a) original narrow-band signal; (b) output signal of the proposed SQE system; (c) original wide-band signal.

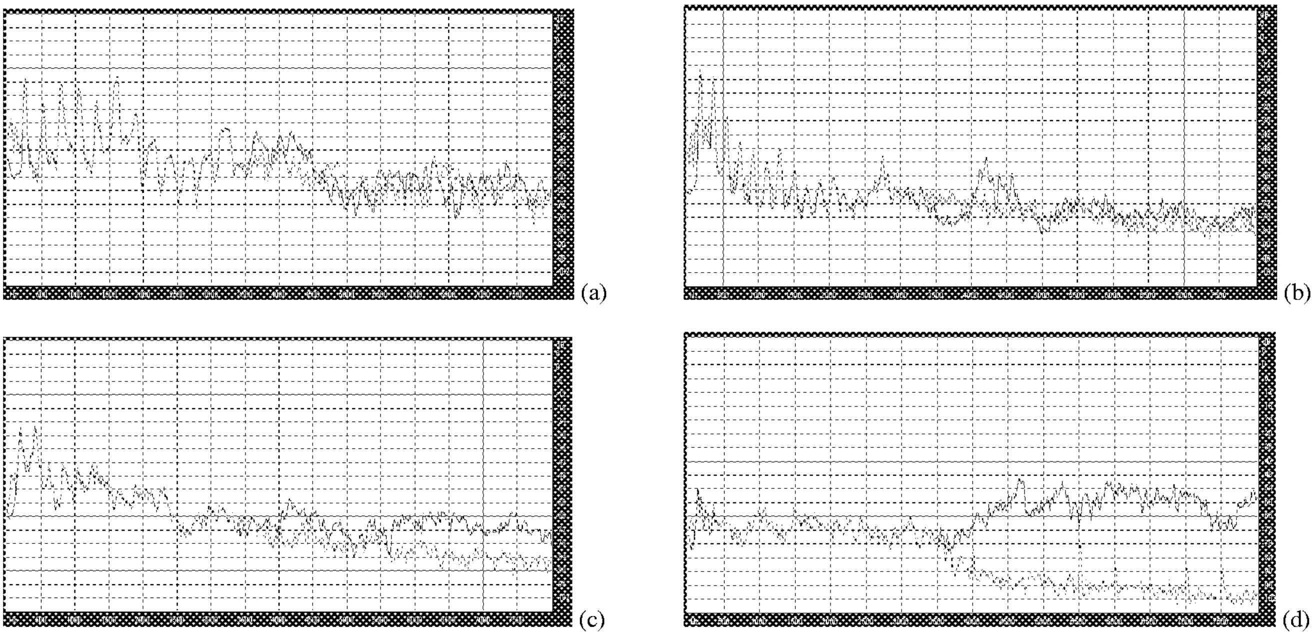

Figure 6. Comparison of the original wide-band spectra of different speech sounds (black solid line) with the corresponding outputs from the SQE system (grey solid line): (a) voiced; (b) nasal; (c) unvoiced stop; (d) unvoiced.