Proc. 13th Annual Conf. Int. Speech Communication Association (Interspeech 2012), Portland, Oregon, 9 - 13 September 2012, Paper No. 689

Real-time Implementation of Multi-band Frequency Compression for Listeners with Moderate Sensorineural Impairment

Nitya Tiwari[1], Prem C. Pandey[2], Pandurangarao N. Kulkarni[3]

[1,2]EE Dept. Indian Institute of Technology Bombay, Mumbai 400076, Maharashtra, India
[3]ECE Dept. Basaveshwar Engineering College, Bagalkot 587102, Karnataka, India
{nitya, pcpandey} @ ee.iitb.ac.in, pnk_bewoor @ yahoo.com

Abstract

Widening of auditory filters in persons with sensorineural hearing impairment leads to increased spectral masking and degraded speech perception. Multi-band frequency compression of the complex spectral samples using pitch-synchronous processing has been reported to increase speech perception by persons with moderate sensorineural loss. It is shown that implementation of multi-band frequency compression using fixed-frame processing along with least-squares error based signal estimation reduces the processing delay and the speech output is perceptually similar to that from pitch-synchronous processing. The processing is implemented on a DSP board based on the 16-bit fixed-point processor TMS320C5515, and real-time operation is achieved using about one-tenth of its computing capacity.

Index Terms: sensorineural hearing loss, multi-band frequency compression, real-time processing

1. Introduction

Sensorineural hearing impairment is generally associated with elevated hearing thresholds, loudness recruitment and reduced dynamic range, and increased temporal and spectral masking leading to degraded speech perception [1]-[3]. In addition to providing frequency-selective gain and automatic gain control, many hearing aids have multi-channel dynamic range compression with settable attack time, release time, number of channels, and compression ratios [3],[4]. For further improving speech perception, several techniques have been reported for reducing the effect of increased intraspeech spectral masking caused by widening of auditory filters [5]-[12].

Binaural dichotic presentation, using a pair of comb filters with complementary magnitude responses for spectral splitting, has been used for persons with moderate bilateral loss [5],[6]. In case of monaural hearing, spectral contrast enhancement [7],[8] may reduce the effects of widened auditory filters. It involves enhancing the perceptually important spectral peaks, thus increasing the contrast between spectral peaks and valleys. However, errors in identifying these peaks may subdue the advantage of spectral contrast enhancement and increase in the dynamic range due to contrast enhancement may adversely affect speech perception by persons with reduced dynamic range. Multi-band frequency compression [9]-[12] is another technique for reducing the effects of intraspeech spectral masking. In this technique, speech energy is presented in relatively narrow bands to avoid masking by adjacent spectral components. The processing involves dividing speech spectrum into analysis bands and compressing the spectral samples in each band towards the band center using a constant compression factor. Arai et al. [9] applied the technique using auditory critical bandwidths on the magnitude spectrum and the complex spectrum was obtained by associating the compressed magnitude spectrum with the original phase spectrum. For decreasing the computation and reducing the processing related artifact, Kulkarni et al. [11] applied the compression on the complex spectrum without calculating the magnitude and phase spectra.

The processing for multi-band frequency compression involves three steps: (i) segmentation and spectral analysis, (ii) spectral modification, and (iii) resynthesis. The effects of different types of segmentation, bandwidths, and frequency mappings for spectral modification have been reported earlier [11]. Three frequency mapping techniques were investigated: (i) sample-to-sample mapping, (ii) spectral sample superimposition, and (iii) spectral segment mapping. Three types of bandwidths were used: (i) fixed bandwidth with number of bands varying from 2 to 18, (ii) 1/3-octave bands, and (iii) bands based on auditory critical bandwidth, over 0 – 5 kHz [13]. Two types of segmentation were investigated. Fixed-frame segmentation used 20 ms analysis frame with 50 % overlap. Pitch-synchronous segmentation used frame length equal to two local pitch periods with an overlap of one pitch period. The pitch period was estimated by detection of glottal closure instants (GCIs) [14]. During the unvoiced segments, the pitch period of the last voiced segment was used. Best results were obtained for auditory critical bandwidth based compression using spectral segment mapping and pitch-synchronous analysis-synthesis. Evaluation by conducting Modified Rhyme Test on eight subjects with moderate sensorineural loss showed best performance for the compression factor of 0.6 [12]. There was a mean increase of 16.5% in recognition scores and a mean decrease of 0.89 s in response time, and the improvements were statistically significant for all the listeners. Thus the results showed that the processing is effective in improving consonant recognition with a reduced load on the perception process and that its use in hearing aids may improve speech perception by persons with moderate sensorineural loss.

For its use in hearing aids, the processing needs to be implemented for real-time operation. Fixed-frame analysis-synthesis is suitable for real-time operation, but it results in perceptible distortion. Pitch-synchronous analysis-synthesis avoids this distortion and results in better scores [11]. But pitch estimation adds to algorithmic and computational delays. To solve this problem, application of the Griffin-Lim method of signal estimation from modified short-time spectrum [15] is investigated. An implementation of the technique for real-time processing on a 16-bit fixed-point DSP board is presented.

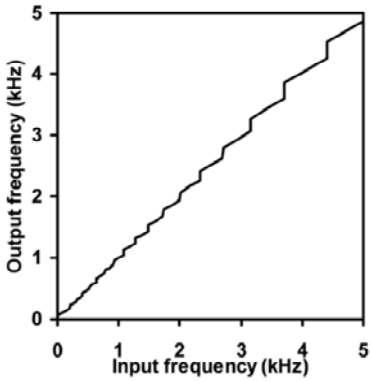

Figure 1: *Frequency mapping with α=0.6 [11].*

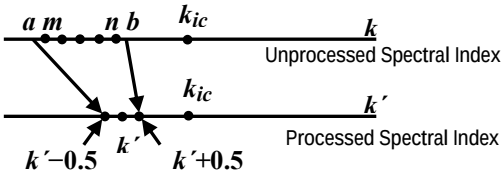

Figure 2: *Spectral segment mapping [11].*

2. Signal processing

In the multi-band frequency compression with fixed-frame processing [11], input signal is segmented using analysis window of fixed length *L* with 50% overlap. Each segment is zero padded, and *N*-point DFT is calculated. Spectral modification is applied on the complex spectral samples using auditory critical bands and compression factor α, with the frequency mapping as shown in Figure 1. The output is obtained by applying overlap-add on *N*-point IDFT of the modified complex spectrum. Spectral modification is carried out using spectral segment mapping as shown in Figure 2. It involves a uniform contribution from all samples of the input spectrum to the compressed spectrum, and it results in no perceptible artifacts [11]. A one-sample interval centered on the output frequency sample is mapped to a corresponding segment of the frequency axis of the input spectrum. The edges *a* and *b* of the input frequency segment for the output spectral sample with frequency index *k′* in the *i*th analysis band with center frequency k_{ic} are given as

$$a = k_{ic} - [(k_{ic} - (k' - 0.5))/\alpha] \quad (1)$$

$$b = a + 1/\alpha \quad (2)$$

The sample in the compressed spectrum is calculated from samples of the input complex spectrum as

$$Y(k') = (m\text{-}a)\,X(m) + \sum_{j=m+1}^{n-1} X(j) + (b-n)\,X(n) \quad (3)$$

where *m* and *n* are the indices of the first and the last spectral samples in [*a*,*b*], respectively. Spectra of 100 ms segments of the processed and unprocessed vowel /*i*/ and broad-band noise for α = 0.6, given in Figure 3, show that the processing concentrates the spectral energy in narrow bands and it does not introduce any spectral tilt.

Speech output from fixed-frame processing has perceptible distortion. Listening test with normal-hearing listeners showed that the scores for the processed speech were higher than that for the unprocessed speech only in the presence of masking noise, indicating that perceptible distortions adversely affected the advantage of frequency compression. Pitch-synchronous processing removed the distortion and resulted in a significant

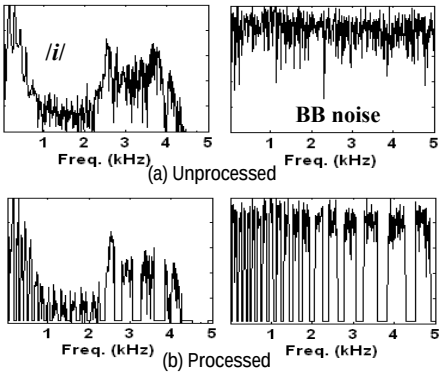

Figure 3: *Spectra of vowel /i/, and broad-band noise.*

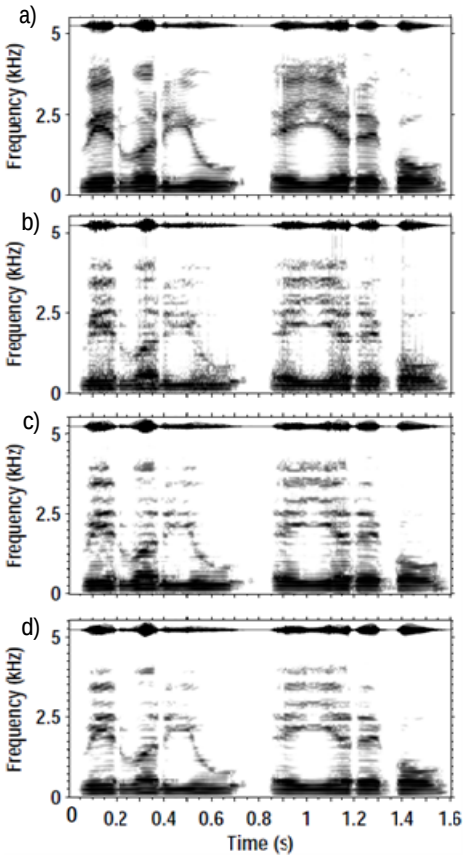

Figure 4*: Wide-band spectrograms of the sentence "where were you a year ago?": (a) unprocessed (b) fixed-frame processed, (c) pitch-synchronous processed, (d) LSEE processed (10 kHz sampling, 26 ms window, N= 1024, α=0.6).*

improvement in speech perception [11],[12]. However, it is not suitable for real-time operation because of the algorithmic and computational delays associated with the detection of glottal closure instants.

In a processing involving modification of short-time Fourier transform (STFT), the resulting spectrum, in general, may not be a valid STFT in the sense that it may not be possible to associate a time-domain sequence with it. In practical terms, it happens because of discontinuities between the signal segments corresponding to the consecutive modified complex spectra. Use of 50% overlap-add in the fixed-frame processing helps in masking the discontinuity, but it leads to an artifact in the form of another superimposed pitch which is

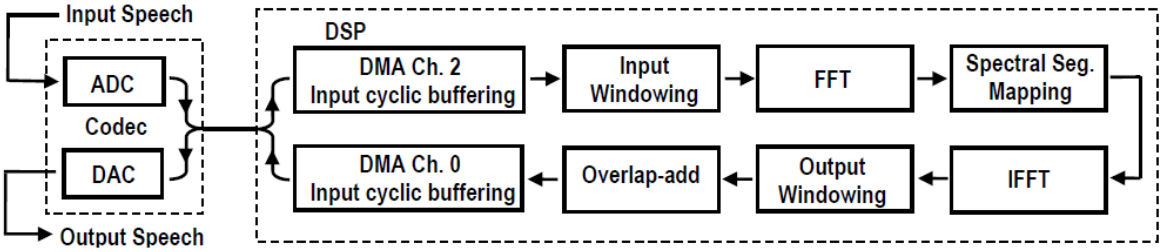

Figure 5: *Block diagram of implementation of multi-band frequency compression on the DSP board*

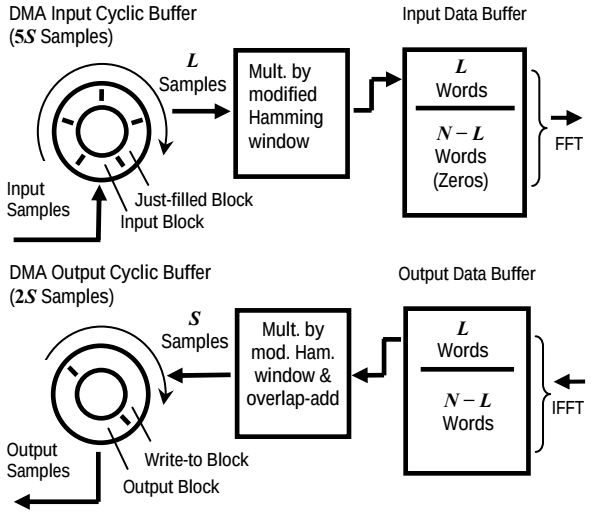

Figure 6: *Data transfer and buffering operations (S = L/4).*

related to the window shift duration. Pitch-synchronous processing with window length of two local pitch periods and one pitch-period overlap avoids this problem.

To avoid the artifact associated with the fixed-frame processing with 50% overlap and the additional algorithmic and computational delay associated with the pitch-synchronous processing, we have used the Griffin-Lim method [15] of signal estimation from modified short-time complex spectrum. The method is based on least squared error estimation (LSEE), i.e., minimizing the mean squared error between the modified STFT and the STFT of the estimated signal. The output signal is resynthesized by overlap-add of the segments obtained as IDFT of the modified complex spectra after multiplication with the analysis window. The window used should meet the requirement that sum of the squares of all the windows is unity, i.e.,

$$\sum_{m=-\infty}^{\infty} w^2(mS - n) = 1 \qquad (4)$$

For window length L and window shift $S = L/4$ corresponding to 75% overlap, this requirement is met by modified Hamming window, given as

$$w(n) = [1/\sqrt{(4p^2 + 2q^2)}][p + q\cos(2\pi(n + 0.5)/L)] \qquad (5)$$

with $p = 0.54$ and $q = -0.46$. Griffin and Lim [15] extended the method to reconstruct a signal from the modified short-time magnitude spectrum by using an iterative technique. Subsequently, other methods for signal estimation from modified short-time magnitude spectrum and suited for real-time processing have been reported [16]. Since the spectral modification using multi-band frequency compression results in complex spectra, it can be easily implemented using LSEE method. It is subsequently referred to as LSEE processing. For a comparison of the pitch-synchronous and LSEE processing, both were implemented using MATLAB. The speech outputs from the two implementations were perceptually similar and the PESQ-MOS was found to be 3.7. Wide-band spectrograms of the unprocessed and processed signal for the sentence "where were you a year ago?" are shown in Figure 4. The spectrograms of processed speech show the concentration of spectral energy into narrower bands. The vertical striations indicate that the harmonic structure is approximately preserved. The processing retains the formant transitions with only slight shift in the formant locations. Processing artifact in the form of discontinuities is observed in the spectrogram of the output from fixed-frame processing, but not in those from pitch-synchronous and LSEE processing. LSEE processing is found to be well suited for non-speech audio also, as it does not require pitch estimation.

3. Real-time implementation

Multi-band frequency compression is implemented for real-time processing on DSP board "eZdsp", based on 16-bit fixed-point processor TI/TMS320C5515 [17]. The processor can be operated at a clock frequency of up to 120 MHz. It has a unified memory space of 16 MB. The on-chip features include 320 KB RAM (with 64 KB dual-access data RAM), 128 KB ROM, four 4-channel DMA controllers, three 32-bit timers, FFT hardware for efficiently computing 8 to 1024-point FFT. A complex number is stored as 4-byte word, with 16-bit real and 16-bit imaginary parts. The board has 4 MB flash for user program and codec TLV320AIC3204 [18] with stereo ADC and DAC with 16/20/24/32-bit quantization and 8 – 192 kHz sampling. Our implementation uses one channel of the stereo codec with 16-bit quantization and 10 kHz sampling. The program was written in C, using TI's 'CCStudio, ver. 4.0' as the development environment. For reducing conversion overheads, the input samples, spectral values, and the processed samples are stored as complex numbers. The imaginary part of input sample is set to zero.

Figure 5 shows the block diagram of the implementation. Codec and DMA are used to acquire and output the speech signal. The data transfer and buffering operations are shown in Figure 6. Signal acquisition uses a 5-block DMA input cyclic buffer, with S-word blocks. A buffer of N words, initialized with zero values, serves as input data buffer. At regular sampling intervals, DMA channel-2 reads the input from ADC and writes to the DMA input cyclic buffer. The output is handled using a 2-block cyclic buffer, with S-word blocks. DMA channel-0 is used to cyclically output to DAC. Pointers keep track of the current input, just-filled input, current output, and write-to output blocks, and these are initialized to 0, 4, 0, and 1, respectively. A DMA interrupt is generated when a block gets filled. All pointers are incremented cyclically. The DMA-mediated reading from ADC and writing to DAC are continued. The samples of the just-filled and the previous

three blocks are copied to the input data buffer and multiplied by modified Hamming window of length L as given in (5). These samples padded with $N-L$ zero-valued samples serve as input to N-point DFT. This method of data handling, results in an efficient realization of 75 % overlap and zero padding.

N-point complex DFT of the input array is stored in an N-word buffer. The first $N/2$ spectral samples of the compressed spectrum are calculated using spectral segment mapping as given in (3), with the other samples remaining zero valued. A look-up table of pre-calculated values of m, n, $m-a$, and $b-n$ for each output spectral index is used. The time domain segment is obtained as the first L samples of the real part of the N-point IDFT of the modified complex spectrum and multiplied by twice the modified Hamming window. The result is stored in the output data buffer. Overlap-add operation uses a buffer of $3S$ samples. The first S samples of the output data buffer are added to the first S samples of the overlap buffer containing the partial results from the previous operation. The resulting samples are written as the processed output to the write-to output block. The next $2S$ samples of the output data buffer and the overlap buffer are added together and copied as the first $2S$ samples of the overlap buffer. The last S samples of the output data buffer are copied as the last S samples of the overlap buffer. For real-time processing, all the operations on the samples in the input data buffer should get completed within S sampling intervals.

4. Results

The processing used sampling frequency of 10 kHz, window length $L = 260$ (i.e. 26 ms), and DFT size $N = 1024$. Informal listening tests showed that the processed output from the DSP board was perceptually similar to the corresponding output from the offline implementation for speech as well as other audio signals. The program operation was tested by progressively decreasing the clock frequency from 120 MHz and it worked satisfactorily down to 20 MHz. Use of $N = 512$ did not result in any perceptible change in the output and permitted satisfactory operation for clock frequency down to 12 MHz, indicating a significant amount of unused computational capacity which may be useful in implementing other processing as needed for a hearing aid. The processing has an algorithmic delay of L samples and computational delay of less than $L/4$ samples. For 26 ms window, the total processing delay was found to be less than 35 ms, which is acceptable for its use in the hearing aids along with lipreading.

The processed output signal was acquired through a PC sound card and the spectrograms matched with those from offline implementation. For compression factors of 0.6–1.0, PESQ-MOS between the offline and real-time speech outputs was 2.5–3.4.

5. Conclusion

It has been shown that implementation of multi-band frequency compression using fixed-frame processing along with the Griffin-Lim method of signal estimation from the modified short-time complex spectra results in a processed output which is perceptually similar to the output from the pitch-synchronous processing. It has been implemented for real-time processing using the 16-bit fixed point processor TMS320C5515, using about one-tenth of its computing capacity. For use of the processing in hearing aids, frequency-selective gain and multi-band dynamic range compression, with the gain and compression ratios settable in accordance with the loss characteristics of the individual listener, also need to be implemented.

6. Acknowledgements

The research is supported by the project “National Program on Perception Engineering”, sponsored by the Department of Information Technology, MCIT, Government of India.

7. References

[1] H. Levitt, J. M. Pickett, and R. A. Houde, Eds., Sensory Aids for the Hearing Impaired. New York: IEEE Press, pp. 3–10, 1980.

[2] B. C. J. Moore, An Introduction to the Psychology of Hearing, London, UK: Academic, 1997, pp 66–107.

[3] J. M. Pickett, The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology. Boston, Massachusetts: Allyn Bacon, 1999, pp 289–323.

[4] H. Dillon, Hearing Aids. New York: Thieme Medical Publisher, 2001.

[5] T. Lunner, S. Arlinger, and J. Hellgren, “8-channel digital filter bank for hearing aid use: preliminary results in monaural, diotic, and dichotic modes,” Scand. Audiol. Suppl., vol. 38, pp. 75–81, 1993.

[6] P. N. Kulkarni, P. C. Pandey, and D. S. Jangamashetti, “Binaural dichotic presentation to reduce the effects of spectral masking in moderate bilateral sensorineural hearing loss”, Int. J. Audiol., vol. 51, no. 4, pp. 334–344, 2012.

[7] T. Baer, B. C. J. Moore, and S. Gatehouse, “Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times”, Int. J. Rehab. Res., vol. 30, no. 1, pp. 49–72, 1993.

[8] J. Yang, F. Luo, and A. Nehorai, “Spectral contrast enhancement: Algorithms and comparisons”, Speech Commun., vol. 39, no. 1–2, pp. 33–46, 2003.

[9] T. Arai, K. Yasu, and N. Hodoshima, “Effective speech processing for various impaired listeners”, in Proc. 18th Int. Congr. Acoust., 2004, Kyoto, Japan, pp. 1389–1392.

[10] K. Yasu, M. Hishitani, T. Arai, and Y. Murahara, “Critical-band based frequency compression for digital hearing aids,” Acoustical Science and Technology, vol. 25, no. 1, pp. 61-63, 2004.

[11] P. N. Kulkarni, P. C. Pandey, and D. S. Jangamashetti, “Multi-band frequency compression for reducing the effects of spectral masking,” Int. J. Speech Tech., vol. 10, no. 4, pp. 219–227, 2009.

[12] P. N. Kulkarni, P. C. Pandey, and D. S. Jangamashetti, “Multi-band frequency compression for improving speech perception by listeners with moderate sensorineural hearing loss,” Speech Commun., vol. 54, no. 3 pp. 341–350, 2012.

[13] E. Zwicker, “Subdivision of the audible frequency range into critical bands (Freqenzgruppen),” J. Acoust. Soc. Am., vol. 33, no. 2, pp. 248, 1961.

[14] D. G. Childers and H. T. Hu, “Speech synthesis by glottal excited linear prediction”, J. Acoust. Soc. Am., vol. 96, no. 4, pp. 2026–2036, 1994.

[15] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time Fourier transform,” IEEE Trans. Acoustics, Speech, Signal Proc., vol. 32, no. 2, pp. 236-243, 1984.

[16] X. Zhu, G. T. Beauregard, and L. L. Wyse, “Real-time signal estimation from modified short-time Fourier transform magnitude spectra”, IEEE Trans. Audio, Speech, Language Process., vol. 15, no. 5, pp. 1645–1653, 2007.

[17] Texas Instruments Inc., TMS320C5515 Fixed-Point Digital Signal Processor. 2011, [online] Available: http://focus.ti.com/lit/ds/symlink/tms320c5515.pdf.

[18] Texas Instruments Inc., TLV320AIC3204 Ultra Low Power Stereo Audio Codec. 2008, [online] Available: http://focus.ti.com/lit/ds/symlink/tlv320aic3204.pdf.