BIMODAL PERCEPTION OF SPECTRUM COMPRESSED SPEECH

Larry D. Paarmann
Department of Electrical Engineering
Wichita State University
Wichita, Kansas

Michael K. Wynne
Department of Otolaryngology
Indiana University School of Medicine
Indianapolis, Indiana

ABSTRACT

In this paper, the results of both normal-hearing, and profoundly hearing-impaired adults, tested with spectrum compressed speech via the modified chirp-z algorithm, with and without visual stimuli, are reported.

Ten normal-hearing adult listeners and three profoundly hearing-impaired adult listeners were asked to identify nonsense syllables presented auditorily and bimodally (audition and vision) via video tape in two conditions: lowpass filtered or unprocessed, and spectrum compressed. The lowpass filtered and spectrum com-pressed speech occupies the same spectrum width of 840 Hz; at 900 Hz and above, the attenuation is at least 60 dB. The spectrum compression is performed by means of a modified chirp-z algorithm, and is described in this paper. The testing results are significant and are reported in this paper.

1. INTRODUCTION

The intended population that the subject of this paper addresses are those who are profoundly hearing impaired, such that conventional hearing aids are of little if any use, but still have significant residual hearing. Such subjects are now being considered for cochlear implants. But the authors of this paper believe that more sophisticated methods of signal processing have not been fully exploited, can significantly assist such subjects, and, of course, offers a non-invasive solution.

Space does not permit a literature review, however, the interested reader is referred elsewhere for a fairly extensive review [1]. More recent literature may also be readily found [2]-[7].

2. MATHEMATICAL BASIS

Let N samples of $X(z)$ be expressed as follows:

$$X(z_k) = \sum_{n=0}^{N-1} x(n)\, z_k^{-n}\ , \quad 0 \le k \le N-1 \quad . \qquad \textbf{(1)}$$

Equation **(1)** may be expressed in matrix notation as follows:

$$\underline{X} = \Phi \underline{x} \quad , \qquad \textbf{(2)}$$

where

$$\underline{X} = [X(z_0),\ X(z_1),\ \cdots,\ X(z_{N-1})]^T \quad ,$$

$$\underline{x} = [x(0),\ x(1),\ \cdots,\ x(N-1)]^T \quad ,$$

and

$$\Phi = \begin{bmatrix} 1 & z_0^{-1} & z_0^{-2} & \cdots & z_0^{-N+1} \\ 1 & z_1^{-1} & z_1^{-2} & \cdots & z_1^{-N+1} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & z_{N-1}^{-1} & z_{N-1}^{-2} & \cdots & z_{N-1}^{-N+1} \end{bmatrix} \quad .$$

If $z_k \ne z_i$ for all $k \ne i$, then Φ will be full rank, and the existence of Φ^{-1} is assured, since Φ is recognized as being a Vandermonde matrix [8]. Then, from **(2)**:

$$\underline{x} = \Phi^{-1} \underline{X} \quad . \qquad \textbf{(3)}$$

That is, the N-point $\underline{x}$ is recoverable from any set of N distinct samples of $X(z)$ by **(3)**. This result has significance here, in that it demonstrates that any algorithm which samples $X(z)$ in any fashion, obtaining N distinct samples from an N-point $\underline{x}$, does not lose information.

Note that the Discrete Fourier Transform (DFT) is a special case of the above:

$$X(e^{j\omega_k}) = X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j\frac{2\pi}{N}nk}\ , \quad 0 \le k \le N-1 \quad ,$$

where

$$z_k = e^{j\omega_k} = e^{j\frac{2\pi}{N}k} \quad .$$

The nonlinearly-spaced sampling used for nonlinear spectrum compression is a modification of the chirp-z algorithm [9],[10]. Let the z-domain samples be expressed as follows:

$$z_k = W(k) = W_o(k)\, e^{j\,\phi(k)} \quad ,$$

where $W_o(k)$ and $\phi(k)$ are both real: $W_o(k)$ specifies the magnitude of the z domain samples and $\phi(k)$ specifies the angle. If $0 \le k \le N-1$, then $W_o(0)$ and $\phi(0)$ will specifiy the beginning sample, and $W_o(N-1)$ and $\phi(N-1)$ will specify the ending sample. The sampling contour will be specified by $W_o(k)$ and $\phi(k)$. Detail may be found elsewwhere [1].

3. SPECTRUM COMPRESSION ALGORITHM

To properly follow temporal changes in speech, window lengths of 10 to 20 *ms* are used. In the spectrum compression algorithm, each windowed frame of input speech data is processed by the algorithm prior to overlap-add reconstruction, resulting in an output speech signal that follows the spectral dynamics of the input speech, but has its spectrum compressed and lowered according to the spectrum compression algortihm.

The spectrum compression algorithm includes concepts from frequency transposition. The only spectral range that is compressed is that which is typically above the range of voiced-speech formants. These high frequencies, typically unvoiced fricatives, are spectrally compressed and transposed to a lower frequency range.

To avoid excessive spectral compression that would make increased demands on the spectral resolution capabilities of the listeners, critical band theory is adhered to. That is, the critical bands across the spectral range of the input speech that is to be spectrally compressed are mapped to the same number of critical bands in the processed speech spectrum.

The approach uses two parallel channels, and results in a compressed spectrum width of 840 Hz. This spectrum width was somewhat arbitrarily chosen, but is appropriate for the intended population. Word recognition ability of such subjects is very poor, as can be seen in the data below, and yet they have sufficient residual auditory function to be effectively assisted by appropriate signal processing.

Refer to Figure 1, where the algorithm is illustrated in block diagram form. The input speech spectrum width is wideband (10.5 kHz), and is sampled at 25 k samples/s with a 16 bit A/D. Channel 1, illustrated in Figure 1, is simply lowpass filtering. The output of Channel 1 has a spectrum width of 840 Hz (60 dB down, or more, for all frequencies above 900 Hz).

Channel 2 begins by windowing the speech data with a Hanning window. The windowed speech data are nonlinearly sampled in the frequency domain for spectrum compression and transposed to a low frequency region. The outputs of Channels 1 and 2, as illustrated in Figure 1, are added together. The frequency ranges are such that a vowel will result in an output from Channel 1 with essentially no output from Channel 2, and an unvoiced fricative will result in an output from Channel 2 with essentially no output from Channel 1. Details of Channel 2 processing are illustrated in Figure 2.

The frequency domain mapping in Channel 2 is illustrated in Figure 3. Four and one half critical bands, extending from 4,400 Hz to 10,500 Hz, are mapped to 300 to 840 Hz. Note that this requires not only frequency transposition, but nonlinear spectrum compression as well. The processing of Channel 2 has been confirmed by sinewave testing, synthetic data testing, and by spectrographic comparisions with real speech data.

4. EVALUATION METHODS

Thirty-three video tapes have been prepared for this study: 11 tapes of unprocessed, wideband speech, 11 tapes of lowpass filtered

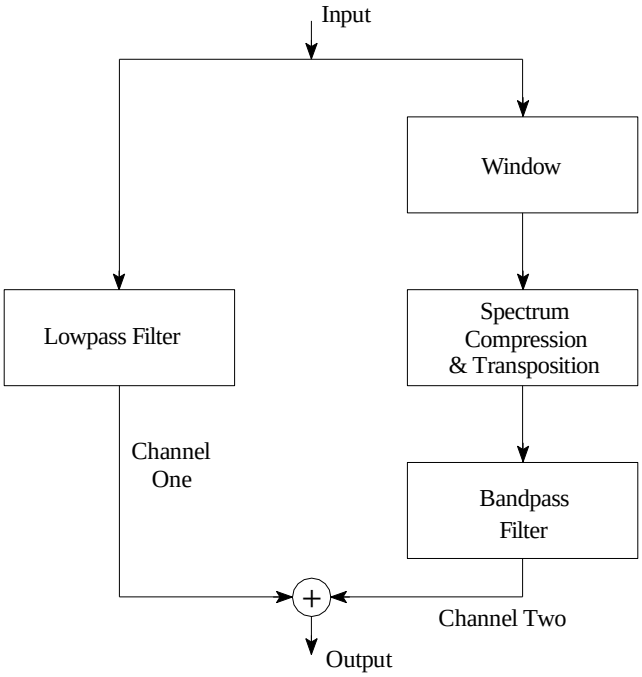


Figure 1. Overall algorithm block diagram.

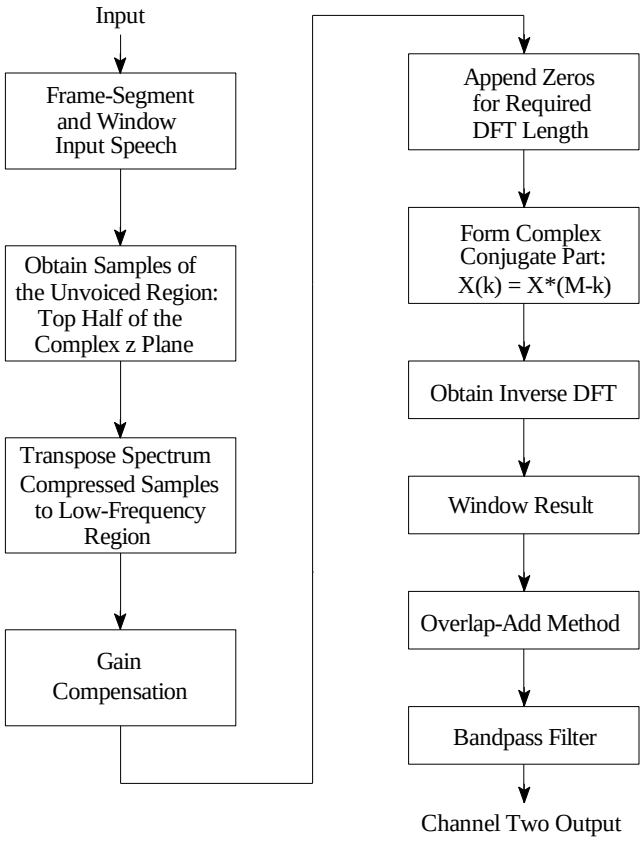


Figure 2. Detailed block diagram of Channel Two.

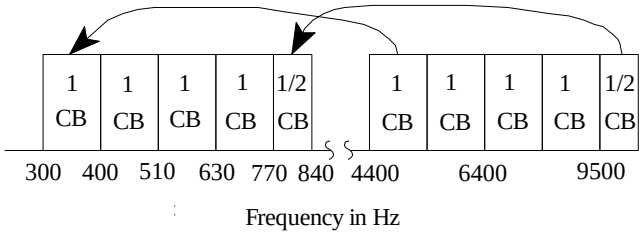


Figure 3. Illustration of the critical band mapping.

speech, and 11 tapes of spectrum compressed speech. Videotape recordings were professionally made of an adult male speaker presenting the stimulus items from 11 pages of nonsense syllables. The speaker was using Standard American English and is extensively trained in phonetics. Each of the 11 pages of nonsense syllables contains 7 to 9 vowel-consonant (VC) or consonant-vowel (CV) stimulus items embedded in 10 to 12 foils. Each foil has a carrier phrase. For example, "You will mark EETH please." The video fades to black for five seconds between each foil. There are ten randomizations of each page. Score sheets have been prepared for each page.

5. RESULTS

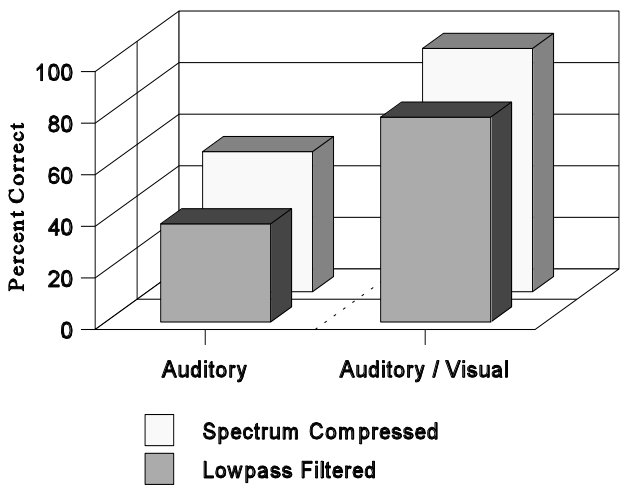
The results of testing with ten normal-hearing subjects are shown in Table 1 and Chart 1. Note that there is a consistent improvement when the subjects are exposed to spectrum compressed speech, compared to lowpass filtered speech occupying the same spectrum width. This is for 102 stimulus items, without any training. Error analysis indicates that these results are significant. Analysis of the confusion matrices suggests that many of the errors made would likely be correctable in practice by the context for spectrum compressed speech. It is anticipated that training would very significantly improve the results for the spectrum compressed speech, since the errors made are understandable and consistent, whereas those for lowpass filtered speech are random; the spectrum compressed consonants are highly distinguishable. The raw data results are shown in Table 2.

The results of testing with three profoundly hearing impaired subjects are shown in Table 3 and Chart 2. The audiogram data for

Table 1: Results of Testing.
Means & Standard Deviations in Percent
of Correct Responses - Normal-Hearing Adults

Listening Mode	Stimulus Condition			
	Lowpass Filtered		Spectrum Compressed	
	Mean	Stan Dev	Mean	Stan Dev
Auditory	38.00	9.89	54.30	8.01
Aud-Vis	79.40	5.29	94.30	3.61

Chart 1: Results of Testing.
Normal-Hearing Adults



the three subjects are summarized in Table 4. Note that the testing was without context, and without training. Also note that one subject showed an improvement from 49% for unprocessed auditory-only syllables to 74% for compressed with visual stimulus syllables. By comparing Table 4 with Table 3 it is clear why Subject No. 1 benefited the most from the spectrum compressed speech, and suggests a subject selection process for screening potential subjects for detailed investigation and testing.

Table 2. Results of Testing, Raw Data.
Percent Correct - Normal-Hearing Adults

Subject Number	Stimulus Condition			
	Lowpass Filtered		Spectrum Compressed	
	Aud Only	Aud-Vis	Aud Only	Aud-Vis
01	41	75	48	91
02	31	79	54	99
03	25	75	53	88
04	58	84	59	98
05	40	86	60	97
06	45	80	65	95
07	33	70	36	88
08	51	88	66	94
09	29	83	50	96
10	27	74	52	97

Table 3. Results of Testing, Raw Data.
Percent Correct - Profoundly Hearing Impaired

Subject Number	Stimulus Condition			
	Unprocessed		Spectrum Compressed	
	Aud Only	Aud-Vis	Aud Only	Aud-Vis
1	49	64	67	74
2	42	49	45	62
3	24	38	20	42

6. CONCLUSIONS

Speech intelligibility is improved when listeners have both auditory and visual information. As place of production cues are provided visually, there is a tendency for listeners to be more sensitive to visual cues when speech signals are degraded.

Since spectrum compressed speech processed by the modified chirp-z algorithm is more intelligible than lowpass filtered speech for normal hearing listeners, and more intelligible than unprocessed speech for profoundly hearing impaired listeners, this signal processing scheme may provide the speech cues for manner production.

7. REFERENCES

1. Paarmann, L.D., Guiher, M.D. and Wynne, M.K., "The Modified Chirp-z Algorithm: Digital Signal Processing of Speech Signals for Nonlinear Frequency Spectrum Compression," *J. Comput. Users Speech Hearing*, vol. 7, pp. 257-281, 1991.

Chart 2: Results of Testing.

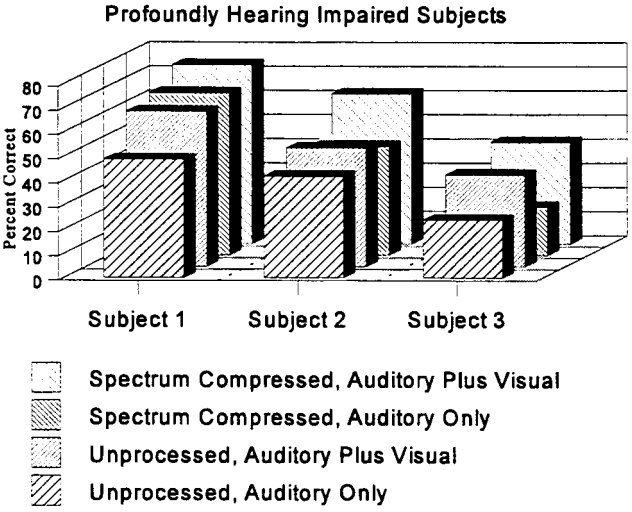


Table 4. Audiogram Data.
Profoundly Hearing Impaired Subjects
Hearing Threshold Level in dB

Subject		Frequency in kHz							
		.25	.5	.75	1	1.5	2	3	4
1	left	30	55	65	85	100	95	110	110
	right	25	55	75	90	100	100	105	110
2	left	70	75	-	100	105	110	-	-
	right	70	75	-	90	110	110	-	-
3	left	60	70	-	75	95	110	105	110
	right	40	65	-	75	85	85	95	100

2. Davis-Penn, W. and Ross, M., "Pediatric Experiences with Frequency Transposing," *Hearing Instrum*., vol. 44, pp. 26-32, 1993.

3. Rosenhouse, J., "A New Transposition Device for the Deaf: a Case Study," *Hearing J*., vol. 43, pp. 20-25, 1990.

4. Levitt, H. and Neuman, A.C., "Evaluation of Orthogonal Polynomial Compression," *J. Acoust. Soc. Amer*., vol. 90, pp. 241-253, 1991.

5. Schulte, C.M., Wynne, M.K. and Paarmann, L.D., "Bimodal Perception of Speech Compressed by a Modified Chirp-z Algorithm," presentation at the annual convention of the *American Academy of Audiology*, Richmond, VA, 1994.

6. Paarmann, L.D., Wynne, M.K., Schulte, C.M. and Bi, J., "Bimodal Perception of Spectrum Compressed Speech," presentation at the 127th meeting of the *Acoustical Society of America*, paper 4pSP15, Cambridge, MA, 1994.

7. Posen, M.P., Reed, C.M. and Braida, L.D., "Intelligibility of Frequency-Lowered Speech Produced by a Channel Vocoder," *J. Rehab. Res. Dev*., vol. 30, pp. 26-38, 1993.

8. Chen, C.-T., *Linear System Theory and Design*, Holt, Rinehart and Winston, New York, NY, 1984.

9. Rabiner, L.R., Schafer, R.W. and Rader, C.M., "The Chirp-z Transform," *IEEE Trans. Audio Electroacoust*., vol. AU-17, pp. 86-92, 1969.

10. Oppenheim, A.V. and Schafer, R.W., *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

APPENDIX: Sound Files

"You will mark EETH please," wideband, lowpass filtered and spectrum compressed: [A628S01.WAV], [A628S02.WAV], [A628S03.WAV]. "You will mark OOSH please," wideband, lowpass filtered and spectrum compressed: [A628S04.WAV], [A628S05.WAV], [A628S06.WAV]. "Your will mark CHA please," wideband, lowpass filtered and spectrum compressed: [A628S07.WAV], [A628S08.WAV], [A628S09.WAV].