

# Speech Coding using Fourier-Bessel Expansion of Speech Signals

Kaliappan Gopalan

Department of Engineering  
Purdue University Calumet  
Hammond, IN 46323, USA  
gopalan@calumet.purdue.edu

**Abstract** – Coding of speech signals using Bessel functions as orthogonal signals in the Fourier-Bessel (FB) expansion has been explored. It has been found that a reasonable quality of speech can be reconstructed using a set of 15 to 30 coefficients in the FB expansion of each frame of speech. At 80 frames per second and eight bits per coefficient, this corresponds to a bit rate of as low as 9600 bits/second when predetermined sequence of coefficients are used. The speech quality and the bit rate increase when higher number or a selected set of coefficients are used. Comparable results in perceptual speech quality and frame-to-frame signal-to-noise were observed for both male and female speakers.

## I. INTRODUCTION

Coding of speech signals for efficient transmission and storage has been typically carried out using the auditory and vocal tract features of the underlying model based on the short-term stationarity of the signals. Although such model-based parametric coders yield low bit-rates, they are relatively complex to implement and susceptible to background noise and transmission errors. Waveform coding methods using linear pulse code modulation (PCM), on the other hand, make no assumption about the signals and are simple and inexpensive to implement; however, PCM bit-rates are generally high and noise is spread throughout the frequency range. Waveform coding using a set of orthogonal functions has the ability to discriminate noise and other unwanted components if these components are orthogonal to the set. This is particularly valid when the orthogonal function set has structural similarity to the signal to be analyzed and coded. For nonstationary signals such as speech, therefore, an aperiodic signal set may be more efficient to use in the representation. Based on this premise, several aperiodic nonsinusoidal functions including exponentially modulated sinusoids and zero order Bessel functions of the first kind have been used for speech analysis with varying degrees of success [1-3]. In the present work we consider the representation and coding of speech waveforms using the first order Bessel functions.

## II. BESSEL FUNCTION REPRESENTATION

Sinusoidal basis functions and the attendant Fourier analysis provide the ability to perform spectral analysis of nonstationary signals based on short-time stationarity of the signals. Unlike the sinusoids, however, Bessel functions of the first kind are quasiperiodic with successive zero-crossing intervals slowly increasing toward  $\pi$ . Because of this similarity of the Bessel functions to short-time speech waveforms, first order Bessel function representation has been considered as more efficient than the Fourier domain description [4-6].

A finite duration signal  $x(t)$  in the interval  $0 \leq t \leq a$  is represented using the first order Bessel functions of the first kind in an infinite Fourier-Bessel (FB) series as [7]

$$x(t) = \sum_{m=1}^{\infty} C_m J_1\left(\frac{x_m t}{a}\right) \quad (1)$$

where  $x_m, m = 1, 2, 3, \dots$  are the roots of  $J_1(t) = 0$ .

Using the orthogonality of the set  $\left\{ J_1\left(\frac{x_m t}{a}\right) \right\}$ , the

FB expansion coefficients are determined from

$$C_m = \frac{2}{a^2 J_0(x_m)^2} \int_0^a t x(t) J_1\left(\frac{x_m t}{a}\right) dt \quad (2)$$

We note that the FB coefficients  $\{C_m\}$  are unique for a given  $x(t)$ , similar to the Fourier series coefficients. Unlike the sinusoidal basis functions in the Fourier series, however, the Bessel functions decay within the range  $a$ , similar to the rise and fall of speech within a pitch interval.

### III. SPEECH WAVEFORM CODING

Because of the redundancy in speech, the infinite series in Eq. (1) can be truncated without appreciable loss of signal integrity. For a speech signal, the truncation can be carried out based on the desirable level of speech quality of the resulting finite series representation or by any objective measure [10]. The series in Eq. (1) may be truncated, for example, simply by using the first N terms. Alternatively, finite dimensionality may be achieved by using a set of selected coefficients in the representation of each frame of speech. We note that the size of the representation and hence the bit rate can be the same in both cases; however, the reconstructed speech quality of the representation may be different because of the difference in the spectral content in each case. Signal reconstruction using as few as 10 coefficients per frame of speech to as high as 150 was studied for the 'first N' term analysis. Since the spectrum of  $s(t) = J_1\left(\frac{x_m t}{a}\right)$  is given by

$$S(\omega) = -j2 \frac{a}{x_m} \frac{\omega}{\sqrt{\left(\frac{x_m}{a}\right)^2 - \omega^2}}, |\omega| < \frac{x_m}{a}, \quad (3)$$

each term,  $C_m J_1\left(\frac{x_m t}{a}\right)$ , in the reconstruction in Eq. (1)

has an approximate bandwidth of  $\omega_B \cong \frac{x_m}{a}$ ; hence, the reconstruction using the first N terms has a maximum bandwidth of  $\omega_{\max} \cong \frac{x_N}{a}$ , where  $x_N$  is the  $N^{\text{th}}$  root of  $J_1(t) = 0$ . Clearly, for small N,  $\omega_{\max}$  is small and hence the reconstructed signal has its spectrum limited to the low end of the original signal spectrum. For high quality speech, therefore, a large number of terms must be used in the reconstruction. For telephone quality of speech with at least 3 kHz bandwidth, for example, N must be approximately 75 for a database of speech sampled at 16,000 samples/s with the range  $a = 200$  samples, or 12.5 ms [4-6]. However, an objective measure such as the long-term signal-to-noise ratio (SNR) may be satisfied even at smaller values of N.

*Other Approximations:* Although a coefficient at a high index m has a wide spectral content from close to zero up to  $\omega_B \cong \frac{x_m}{a}$ , energy at lower frequencies is significantly less than that at  $\omega$  slightly below  $x_m/a$ . Hence, each term in the reconstruction (Eq. 2) may be approximately considered to have a bandpass spectrum with the energy of the spectral components increasing as the frequency increases toward  $x_m/a$ . As a result, the number of terms in the representation and hence the bit rate, can be reduced by selecting terms corresponding to both high and low indices

in Eq. (1). This selection can be made based on speakers and/or speech such that a wide range of frequencies are available in the finite term reconstruction. Alternatively, ordered magnitudes of the coefficients can be used to cover a wider spectrum using fewer coefficients. In this 'selected N' term analysis, 10 to 20 largest magnitude coefficients ( $|C_m|$ ) may be selected out of the first 100 coefficients in each frame. Since large magnitudes are associated with high spectral energies, this choice provides a wider band of spectral energy and a better quality of speech than that resulting from the 'first N' term choice or an arbitrary choice of the same number of coefficients. Sorted coefficients, while yielding higher speech quality, require higher bit rate because of the need to transmit the set of the selected coefficients in each frame.

### IV. EXPERIMENTAL RESULTS

Utterances from the TIMIT database, available as 16-bit integers at 16,000 samples/s, were considered in the experiment. An utterance was segmented into 200 samples/frame (i.e., 12.5 ms frames) with 100-sample overlap, and the first 150 coefficients in the FB expansion of each frame were calculated as given by Eq. (2). The coefficients for an entire utterance were normalized and encoded for transmission. Reconstruction of speech from selected coded coefficients was carried out in accordance with Eq. (1).

In the first case, 10 consecutive coefficients with starting index from 1 to 10 were quantized to 8 bits for reconstruction. At the high index of 10 – when starting index is 1 – a bandwidth of only 410 Hz results for the reconstructed speech. This low bandwidth does not include even the first formant in many cases. Therefore, the starting index was changed to 5 with a final index of 24. This raised the bandwidth to 970 Hz. Although this is still low to cover the first three formants, message conveyed by the test utterance can be perceived with difficulty. Figs. 1 and 2 show the results of reconstruction of a male utterance using 20 coefficients starting at index 5.

Total energy ratio (ER), defined by 
$$ER = \left( \frac{\text{Energy in Original Speech}}{\text{Energy in Recon. Speech}} \right)$$
, was used as a rough measure of comparison between the original and reconstructed signals. With a value of 0.4509 (or original signal energy of approximately 45 percent of reconstructed signal energy) for the case using indices 5 to 24, the reconstructed speech appears to have more total energy than the original speech. While this value does not indicate how comparable the reconstructed utterance quality is compared to the original, it relates the normalized energy in reconstruction. Extremely large or small ratios of ER, clearly, indicate poor approximation to the original speech. As an alternative measure of reconstruction quality, the frame-wise SNR, given by

$$\text{SNR} = \log_{10} \left( \frac{\text{Frame Power in Original Speech}}{\text{Frame Power in Recon. Speech}} \right)$$

was obtained. Frame SNR plot shown in the bottom trace of Fig. 1 – showing a value of close to 0 dB in most frames – demonstrates the closeness of the two signals. The spectrogram of the reconstructed signal given in the bottom trace of Fig. 2 shows the frequency range covered and the trajectory of the first formant. With significant energies in the second and third formants of the original utterance (top trace of Fig. 2), the low bandwidth clearly represents a poor approximation to the original signal. Still, if the utterance is mostly unvoiced, as is the case for the chosen male utterance, a reasonable quality of speech, based on perception, can be obtained from the reconstruction. The bit rate for this corresponds to (20 coefs.)\*(8 bits/coeft.)\*(1/12.5 ms) = 12,800 bits/s.

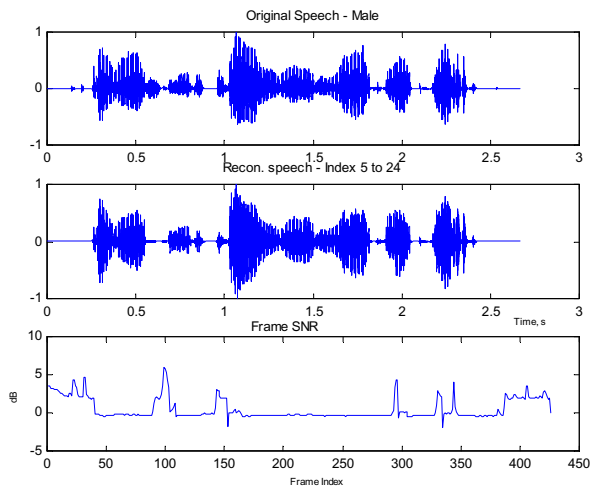


Fig. 1 Original speech (top), reconstructed speech using coefficients 5 to 24 (middle), and the frame SNR (bottom).

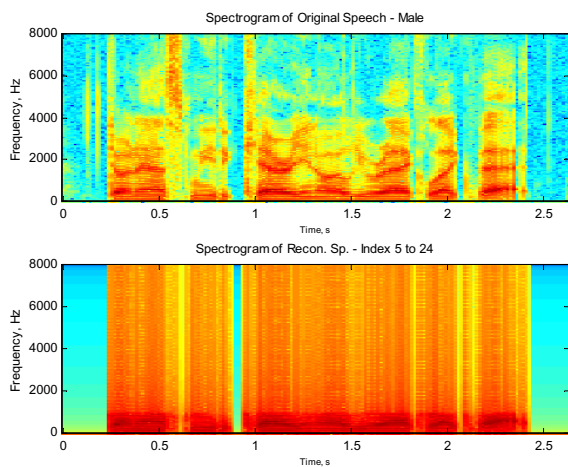


Fig. 2 Spectrogram original speech (top), and reconstructed speech using coefficients 5 to 24

Quantizing the coefficients to 16 bits yields somewhat better speech quality at a cost of twice the bit rate. Still,

the lack of energy in the spectrogram above 970 Hz results in poor speech quality.

A higher bandwidth of up to 3000 Hz, say, and hence improved speech quality, may be obtained using up to the first 75 FB coefficients. In order to preserve the pitch information, however, the lower index cannot be above 5, corresponding to about 210 Hz. Thus the bit rate directly affects the speech quality. Fig. 3 shows the spectrogram of reconstructed speech using coefficients 5 to 75. The frame SNR trace shows that the original signal power is higher than that in the reconstructed signal in many frames. The phonemes in these frames correspond to fricatives carrying energy in frequencies going up to 8000 Hz, as seen from the spectrogram of the original signal in the top trace of Fig. 2. The low bandwidth of 3010 Hz covered by  $x_{75}$  is clearly not sufficient to replicate the original signal. Despite the difference in frame energy, the speech quality is very good. The bit rate, however, has risen to 45,440 bits/s, which is unacceptable for efficient transmission or storage.

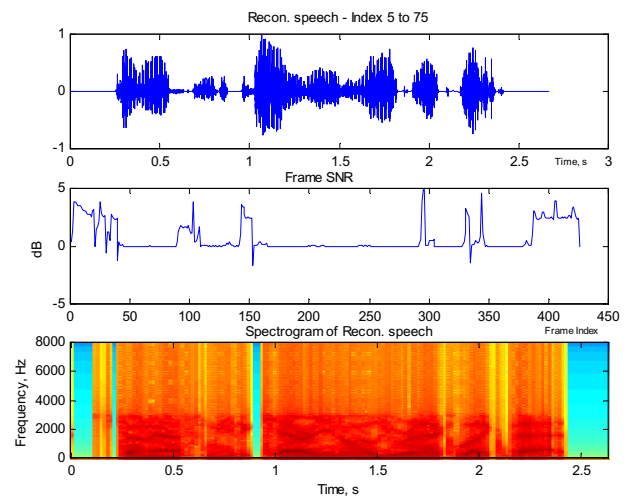


Fig. 3 Reconstruction using consecutive coefficients from index 5 to index 75

Instead of a large number of successive coefficients for obtaining higher spectral band, a reasonable number of coefficients spread across a wider range may be used to achieve better speech quality. The choice of coefficients, however, depends strongly on the type of utterance. If the choice does not include those coefficients corresponding to strong formant frequencies or their neighborhoods, then the perceptual quality will suffer. By selectively using a set of coefficients that covers a wider range in frequency than a set of consecutive coefficients, in general, a better quality can be obtained. Since the same coefficient indices are used in each frame, transmission rate is not unduly increased. Fig. 4 shows the results using an arbitrary choice of indices,

[5:1:20 23:2:70]. We note that with 40 coefficients, the spectrogram of reconstructed speech includes the formant trajectories similar to those using the larger number of coefficients given in Fig. 3. Speech quality, consequently, is about the same as that obtained using indices 5:70. We note here again the difference in energy in frames corresponding to unvoiced areas. With good perceptual quality of the reconstructed speech, the frame energy difference remains the same. The bit rate now is lowered from 44,800 bits/s to 25,600 bits/s. Clearly, with a careful choice of indices, the bit rate can be reduced further without adversely affecting the message quality. The drawback, however, is that the indices of coefficients strongly depends on the spectral content of speech to be transmitted.

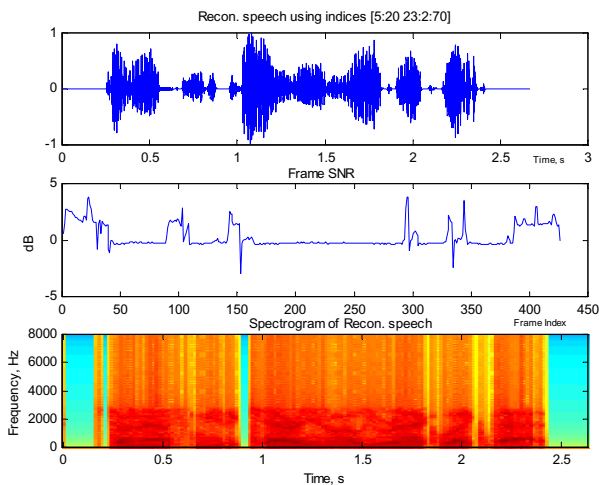


Fig. 4 Reconstruction using an arbitrary set of coefficients

Inasmuch as the spectral energy is concentrated at the formants, FB coefficients reflecting large energy with large magnitudes must be chosen to represent the formants. By using a sufficient number of coefficients with large magnitudes, in general, a wide spectrum that includes the formants can be covered. Although fricatives with low energies at high frequencies may still be not represented adequately, better quality of speech when compared with others choices of coefficients can be obtained. As an example, 20 to 30 coefficients with the largest magnitudes in each frame were seen to approximate the original spectrum well. Fig. 5 gives the reconstructed speech, frame SNR and spectrogram corresponding to 30 largest coefficients. We note that the frame SNR is the smallest of all the previous reconstructions because of the representation of at least some of the high frequency components. The spectrogram reflects the representation of formants and their neighborhoods – high energy regions – as evidenced by the distinct formant trajectories. While the quality of reconstructed speech is much higher than that from any of the previous reconstructions, the bit rate – at  $(30*8 + 30*8)*80 = 38,400$  bits/s – is comparable to rates

in wideband coding. (This assumes that the coefficient indices are coded as unsigned 8 bit integers to cover up to 150, the highest index.) Also, sorting of coefficients in each frame increases the computational complexity. The trade off between quality of speech, transmission rate and complexity is evident from the three choices.

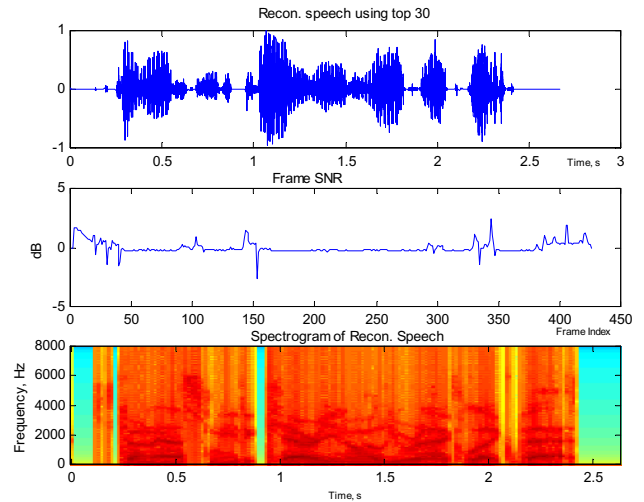


Fig. 5 Reconstruction using the largest 30 coefficients in each frame

The results of the experiments with the choices of selecting coefficients for transmission are summarized in Table I. Clearly, the overall energy ratio ER does not offer much information about the reconstructed speech quality. (The normalized reconstructed signal of the same duration was considered noise in this case.) A somewhat better measure is the frame-to-frame noise shown in Figs. 1, 3, 4 and 5. Even this measure is inadequate in comparing speech quality because of its inability to distinguish in the spectral domain. Elaborate speech quality measures such as in [10] may be used for better comparison in the absence of listening tests.

TABLE I  
COMPARISON OF COEFFICIENT SELECTION FOR TRANSMISSION

Measure → Coefft. choice↓	Bit rate, bits/s	Total Energy Ratio	Speech quality
20 Sequential coeffs	12,800	0.4509	Reasonable
70 Sequential coeffs	44,800	0.8776	Excellent
40 Selected coeffs.	25,600*	0.5064	Reasonable to good
30 Peak coeffs.	38,400	0.5804	Excellent

\* Transmission of selected coefficient indices – to be sent only once – is not included

## V. SECURE TRANSMISSION AND MODIFICATION OF SPEECH

Secure transmission of speech can be achieved as an extension of the 'selected N' term analysis. Coding of a scrambled set of selected FB expansion coefficients with low and high indices, for example, can be used in the transmission and reconstruction of speech. Alternatively, for high quality speech, an ordered set of coefficients and their indices can be transmitted. In this application, speech is processed similar to the 'selected N' waveform coding application with N selected on the basis of the desired quality of reconstructed speech.

Speech modification for voice disguising purposes can be carried out using 'selected N' term analysis in Eq. (1). Unlike in the coding application, however, the choice of coefficients may depend on the spectral content of the utterance. Elimination of a coefficient  $C_m$  in the reconstruction, in general, removes the spectrum of the original signal significantly in the neighborhood of  $x_m/a$ . Hence, in applications requiring emphasizing or deemphasizing of certain range of frequencies, the corresponding coefficients can be modified. As an example, speech modification using native Matlab representation of 64 bits for each altered coefficient was demonstrated to conceal the identity of speakers [9]. With 16 or 32 bits, the modified speech quality may be affected. This can be compensated by using a slightly larger number of coefficients.

## VI. CONCLUSION

Experimental results of coding of speech waveforms using the aperiodic set of Bessel functions of the first order have been presented. These results demonstrate that the FB expansion coding – at 8 bits/coefficient and 12,800 bits/s – gives better speech quality at a lower bit rate than toll speech at 16,000 bits/s. Other applications including secure transmission and speech modification are possible with the choice of coefficients presented.

## VII. REFERENCES

- [1] L. Dolansky, "Choice of Base Signals in Speech Signal Analysis," IRE Trans. Audio, Vol. AU-8, pp. 221-229, 1960.
- [2] H.J. Manley, "Analysis-Synthesis of Connected Speech in terms of Orthogonalized Exponentially Damped Sinusoids," J. Acoust. Soc. Amer., Vol. 35, pp. 464-474, April 1963.
- [3] C.S. Chen, K. Gopalan and P. Mitra, "Speech Signal Analysis and Synthesis via Fourier-Bessel Representation," Proc. ICASSP, Tampa, FL, March 1985, pp. 497-500.
- [4] K. Gopalan, "Speaker Identification using Bessel Function Expansion of Speech Signals," Final Report for Summer Faculty Research Program, Armstrong Laboratory, AFOSR, August 1993.
- [5] K. Gopalan and T. R. Anderson, "Speech Processing using Bessel Functions, Proc. Symp. Intelligent Systems in Communications and Power, Mayaguez, PR, Feb. 1994, pp. 255-259.
- [6] K. Gopalan, T. R. Anderson and E.J. Cupples, "A Comparison of Speaker Identification Results using Features based on Cepstrum and Fourier-Bessel Expansion," IEEE Trans. Speech and Audio Processing, Vol. 7, No.3, pp. 289-294, May 1999.
- [7] I.H.Sneddon, *Fourie Transforms*, New York: McGraw-Hill, 1951.
- [8] A.S. Spanias, "Speech Coding: A Tutorial Review," Proc. IEEE, Vol. 82, No. 10, pp. 1541-1582, Oct. 1994.
- [9] K. Gopalan, "Speech Modification by Selective Fourier-Bessel Series Expansion of Speech Signals," Proc. of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, BC, Canada, Aug. 1999.
- [10] W. Yang, M. Benbouchta and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," ICASSP, Vol. 1, pp. 541-544, Seattle, 1998.