Modeling of Speech Signals Based on Bessel-like Orthogonal Transform

Giorgio Biagetti, Paolo Crippa, Claudio Turchetti

Dipartimento di Elettronica, Intelligenza Artificiale e Telecomunicazioni
Università Politecnica delle Marche, Ancona, Italy
{g.biagetti,pcrippa,turchetti}@deit.univpm.it

Abstract

In this paper a novel modeling technique for speech signals, based on the source-filter model of speech production and on orthogonal transform theory, is presented. The proposed approach models the impulse response of such filter, by projection onto a basis of damped Bessel functions, which have been chosen for their similarity to the signal to be modeled. In such a way an orthogonal transform pair is defined which provides a simple and effective methodology for the extraction of model parameters, and its effectiveness in the case of voiced speech has been demonstrated by synthesizing natural sounding speech signals with the aid of only a few extracted parameters.

Index Terms: speech analysis, impulse response modeling, Bessel functions, orthogonal transforms.

1. Introduction

Speech coding has become a fundamental technology for voice communication over digital networks [1], as new applications such as Voice over IP (VoIP) or mobile communications are gaining popularity over the traditional Plain Old Telephone Service (POTS) carried onto the public switched telephone network.

At the basis of every speech coding technology there is the assumption that speech production and/or perception can be described by appropriate models. One of the first proposed and still widely used model of speech production is the source-filter model, in which the speech signal is considered as the result of a time-varying filter applied to a suitable excitation signal. The filter is supposed to model the effect of speech articulation, while the excitation signal models the acoustic effect of the glottal air flow.

Sinusoidal coding [2] exploits the short-time quasi-stationarity of speech signals to model the output of the filter as a sum of sinusoids that derive from the Fourier expansion of the excitation and are altered by the transfer function of the filter. This approach was very successful and has led to the development of a number of related techniques to capture the most important features of speech signals [3, 4]. Moreover, in [5] and [6] has been noted that, because of the non-periodicity of speech signals, the use of Bessel functions could lead to better results in speech modeling.

In this paper we present a novel speech model that, laying its roots on the source-filter model, uses damped Bessel functions to model the impulse response of the filter, instead of the signal, and a simple pulse train to model the excitation for voiced sounds. Its effectiveness stems from the fact that damped Bessel functions bear a closer resemblance to an impulse response than to an arbitrary portion of the speech signal. A description of a possible modeling technique for voiced sounds and experimental results to support the proposed model will be given next.

2. Orthogonal Transform Pair

The Fourier-Bessel transform employed in [6] attempts to reconstruct a portion (frame) of the speech waveform using Bessel functions, in much the same way as a standard Fourier transform does using sinusoids. The coding gain results from the fact that fewer basis functions are needed to accurately approximate the original speech, because of their greater similarity to the signal to be coded.

In our approach the transform is not applied to the signal itself but it is used to model the filter impulse response $h(t)$. Moreover, a damping factor has been applied to the Bessel functions used as the basis in order to better model the decaying impulse response.

Let $y(t,\lambda)$ be a family of damped Bessel functions of the form:

$$y(t,\lambda) = \mathrm{e}^{-d\,t/2}\sqrt{t} \cdot \left[c_1\, J_\nu(t\sqrt{\lambda - d^2/4}) + c_2\, Y_\nu(t\sqrt{\lambda - d^2/4})\right] \tag{1}$$

where $J_\nu(\cdot)$ and $Y_\nu(\cdot)$ are the first- and second-kind Bessel functions of order ν, respectively, d is a parameter controlling the amount of damping, and c_1, c_2 are real-valued constants. It can be shown that, under wide conditions on $h(t)$, an orthogonal transform pair given by:

$$\mathcal{H}(\lambda) = T\,[h(t)] = \frac{dk(\lambda)}{d\lambda}\int_0^{+\infty} h(t)y(t,\lambda)s^{-2}(t)\,dt \tag{2}$$

$$h(t) = T^{-1}\,[\mathcal{H}(\lambda)] = \frac{1}{\pi}\int_0^{+\infty} y(t,\lambda)\mathcal{H}(\lambda)\,d\lambda \tag{3}$$

can be derived. Here $s(t)$ and $k(\lambda)$ are appropriate normalization functions that depend on the choice of the family defined in (1).

An example of the application of this transform pair is shown in Fig. 1(a)–(b), where a frame extracted from a voiced speech signal and its Bessel transform, respectively, are displayed.

2.1. Convolution properties

Although the standard convolution property, i.e. $T[f \otimes g] = T[f]\,T[g]$, does not hold for the transform pair (2)–(3), an interesting, albeit slightly different property can be derived for a complex version of these transforms.

Let's take the so-called Gabor analytical signal $\hat{\mathcal{H}}(\lambda)$, derived from (2) by means of a Hilbert transform performed in the frequency domain $\omega = \sqrt{\lambda}$, that is:

$$\hat{\mathcal{H}}(\omega^2) = \mathcal{H}(\omega^2) + i\,H[\mathcal{H}(\omega^2)] \tag{4}$$

where $H[\cdot]$ denotes the Hilbert transform. $\hat{\mathcal{H}}(\lambda)$ will be called the *complex Bessel spectrum*. It is easy to verify that, for the time-shifted version:

$$h_\tau(t) = h(t-\tau) \tag{5}$$

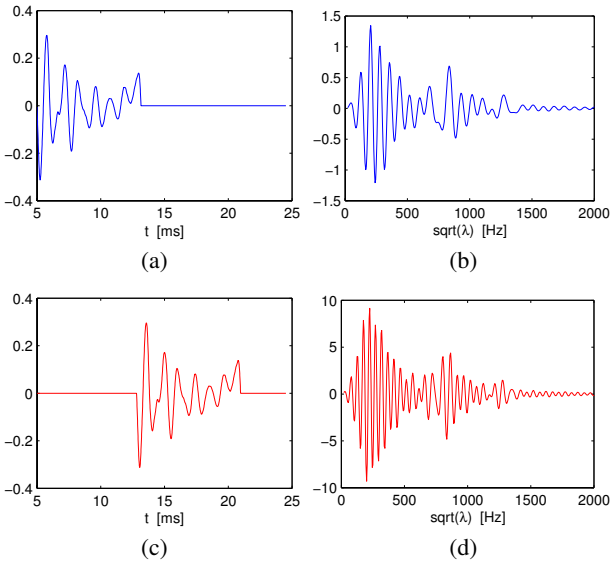


Figure 1: *Examples of Bessel transforms: (a) and (c) are the same frame extracted from a speech signal, differently shifted in time; (b) and (d) are their respective Bessel spectra.*

of the original signal, whose complex Bessel spectrum is $\hat{\mathcal{H}}_\tau(\lambda)$, and that can be seen in Fig. 1(c)–(d), the following remarkable property:

$$\hat{\mathcal{H}}_\tau(\lambda) \approx \hat{\mathcal{H}}(\lambda)\, \mathrm{e}^{(d/2+i\sqrt{\lambda})\tau} \tag{6}$$

holds. This means that, as shown in Fig. 2, the amplitude of the complex Bessel spectrum does not change shape with time shifts, since (6) implies that $|\hat{\mathcal{H}}_\tau(\lambda)| \propto |\hat{\mathcal{H}}(\lambda)|$, with a proportionality factor of $\mathrm{e}^{(d/2)\tau}$.

3. Voiced Speech Model

Let $x(t)$ be a speech signal fragment corresponding to a voiced sound. According to the source-filter model of speech production, $x(t)$ can be thought of as the convolution of an excitation signal $e(t)$ and the vocal-tract impulse response $h(t)$:

$$x(t) = \int_{-\infty}^{t} h(t-\tau)\, e(\tau)\, d\tau\ . \tag{7}$$

Several methods are known to extract an estimate of the excitation signal from the speech signal itself. Most of these try to extract data related to the real glottal air flow, but for our intents a simple pitch-tracking technique will suffice, and thus the excitation signal can be modeled as an impulse train:

$$e(t) = \sum_{j=1}^{M} \delta(t - t_j) \tag{8}$$

where t_j are a rough estimate of the M glottal closure instants, detected by tracking the zero crossings of the speech signal narrowband filtered around the pitch frequency, as discussed later.

By virtue of (6), the complex Bessel spectrum of $x(t)$ can then be written as:

$$\hat{\mathcal{X}}(\lambda) \approx \hat{\mathcal{H}}(\lambda) \sum_{j=1}^{M} \mathrm{e}^{(d/2+i\sqrt{\lambda})t_j} = \hat{\mathcal{H}}(\lambda)\, \mathcal{E}(\lambda) \tag{9}$$

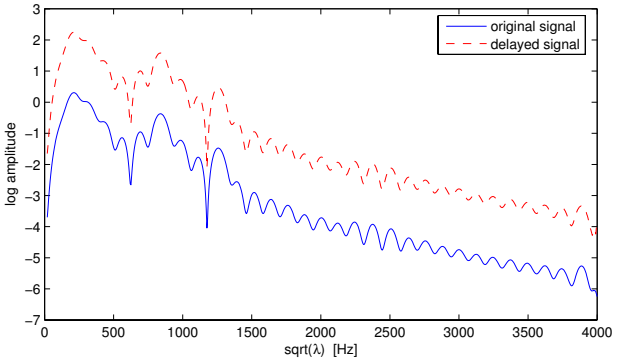


Figure 2: *Demonstration of the convolution property of the Bessel transform: log-amplitude of the Gabor signal derived from the Bessel spectra of Fig. 1(b),(d).*

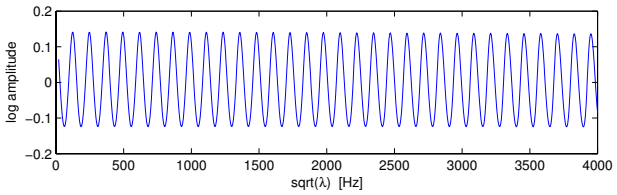


Figure 3: *Amplitude of $\mathcal{E}(\lambda)$ as derived from the extracted pulse train used to model the excitation signal.*

where $\mathcal{E}(\lambda)$ only depends on the excitation signal $e(t)$. In particular, because of the exponential damping, if $h(t)$ is causal and we are only interested in a single speech frame, just a few past ($t_j < 0$) excitation pulses are relevant. An example of a typical signal $\mathcal{E}(\lambda)$, obtained with $M = 5$, is given in Fig. 3.

Finally, the impulse response $h(t)$ can be modeled as a summation of N basis functions derived from the orthogonal transform pair (2)–(3):

$$h(t) = \sum_{k=1}^{N} a_k\, y(t - \tau_k, \lambda_k) \tag{10}$$

where a_k, τ_k, and λ_k are parameters (representing amplitude, phase, and frequency, respectively) that will be extracted from a single speech frame. Strictly speaking, the parameters τ_k would not be necessary, since they do not appear in (3), but they have been included to compensate for the quantization of the others so as to maintain the continuity of the signal phase. Because of their specific use, it is worth noticing that in a typical speech coding application they need not be coded at all, as they can be reconstructed at the decoder.

4. Excitation Model Extraction

To obtain the excitation pulse positions in (8), a very simple algorithm has been employed. The pitch frequency is first estimated by means of Fourier analysis performed on the speech signal. Pitch harmonics are located in the frequency domain and the distances between these are statistically analyzed to compute their mode, which is used to help the identification of the correct peak at the fundamental frequency, even in the presence of spurious neighbouring spectral peaks.

Once the pitch frequency has been estimated, a tight band-pass filter can be designed around such frequency and the zero-crossings of the resulting signal used to derive a first estimate of t_j. Finally, a high-pass filter is used to remove DC offset and low-frequency noise. The parameters t_j are thus refined by aligning them to the closest zero-crossing of the high-pass filtered version of the signal.

5. Impulse Response Model Extraction

In order to estimate $h(t)$, a two-phase approach is used. In the first step a rough approximation to the amplitude of the complex Bessel spectrum $|\hat{\mathcal{H}}(\lambda)|$ is attempted by using (9). At this stage, it may not be convenient to use exactly the value of $\mathcal{E}(\lambda)$ as derived from pitch tracking, because the approximations made could lead to the introduction of spurious peaks into $\mathcal{H}(\lambda)$. Instead, following a procedure somewhat similar to cepstrum-based deconvolution [7], the envelope of the log-amplitude spectrum is taken by interpolating between local maxima, so as to remove the oscillations due to $\mathcal{E}(\lambda)$. In the second step, a peak search algorithm is used to locate the most significant peaks in the complex Bessel spectrum. This is done with the goal of approximating $h(t)$ using only one basis function per peak, by adjusting the parameters in (10).

In detail, the steps involved can be so summarized:

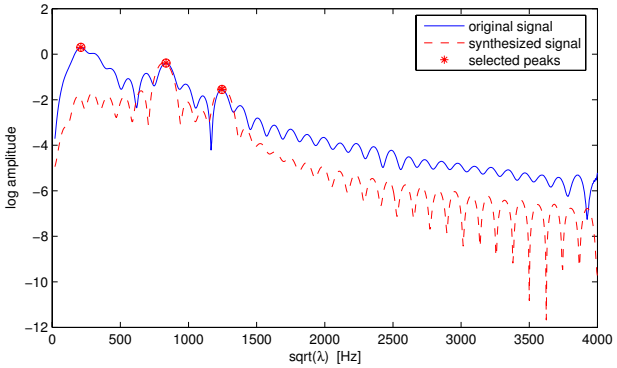
- a pitch-tracking algorithm is used to perform a pitch-synchronous segmentation into speech frames.
- the speech frame $f(t)$ from which we want to extract the impulse response is considered, and its complex Bessel spectrum $\hat{\mathcal{F}}(\lambda)$ is computed.
- a peak search algorithm is used to search the envelope for relevant components to be used as initial guesses for λ_k.
- after the previous step, a tentative set of basis functions to be used to model the impulse response is available. This set is refined using the assumption that if the previous frames had similar spectra and the starting guess was good enough, the selected basis functions convolved with the estimated excitation signal would yield a close approximation to the speech frame under consideration.
- the parameters τ_k and a_k are finally fitted to model the current frame so as to minimize $\|x(t) - f(t)\|$.

Examples of the application of this algorithm will be given in the next section.

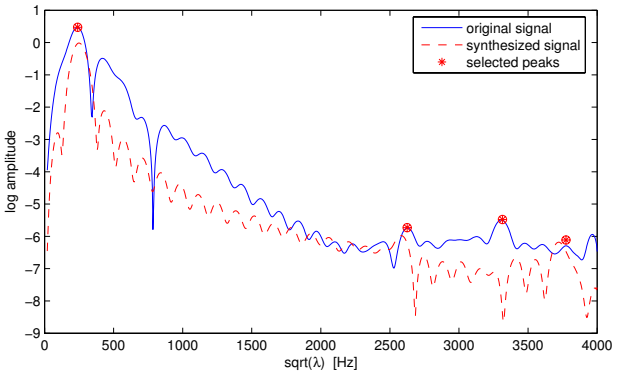
6. Application Examples

In order to test the effectiveness of the proposed technique, the five main Italian vowels (/a/, /e/, /i/, /o/, and /u/) have been modeled, and the reconstructed sounds compared to the original.

Figure 4(a) shows the amplitude of the complex Bessel spectrum of a frame extracted from the original signal corresponding to the vowel /a/ (solid line), with circles marking the frequency position of the basis functions selected by the peak-selection algorithm to model the impulse response. The algorithm was programmed to extract only the four most significant components, i.e. N has been set to 4 in (10). The basis functions so selected have been combined and the result convolved with an impulse train having the same period as the frame being analized. The complex Bessel spectrum amplitude of the resulting signal is also shown in the same figure (dashed line). It is apparent that, although there are some distortions, the main features of the spectrum are preserved. Another example, referred to a more problematic vowel, is shown in Fig. 4(b).



(a)



(b)

Figure 4: *(a) Amplitude of the complex Bessel spectrum of a frame of the Italian vowel* /a/ *(continuous line), and of the synthesized version (dashed line). Circles show the positions of the peaks selected by the peak-selection algorithm. (b) original and synthesized spectra as in (a) but for the Italian vowel* /u/.

The extremely low number of parameters needed to model these voiced sounds suggests that this model could be advantageously used in very-low-bit-rate speech coding applications. To this end, some preliminary informal listening tests that we performed deemed the synthesized speech produced by the proposed model to be of comparable, and, often, superior quality to that of many linear-prediction based vocoders commonly available in the 1 to 3 kb/s bit-rate range, such has Code Excited Linear Prediction (CELP) or Mixed-Excitation Linear Prediction (MELP).

A comparison of original and synthesized speech signals is reported in Fig. 5, where the estimated impulse response, i.e., the selected combination of basis functions weighted with appropriate amplitude coefficients, and a portion of the original signal in the time domain along with a time-aligned reconstruction obtained by convolving the impulse response and the estimated excitation pulse train, is shown for all the vowels used in the test. The impulse response is estimated for a duration equal to five pitch periods, which, as it is apparent from the left column of Fig. 5, are more than enough to make it possible to neglect tails during the reconstruction process. Again, despite the differences in the time-

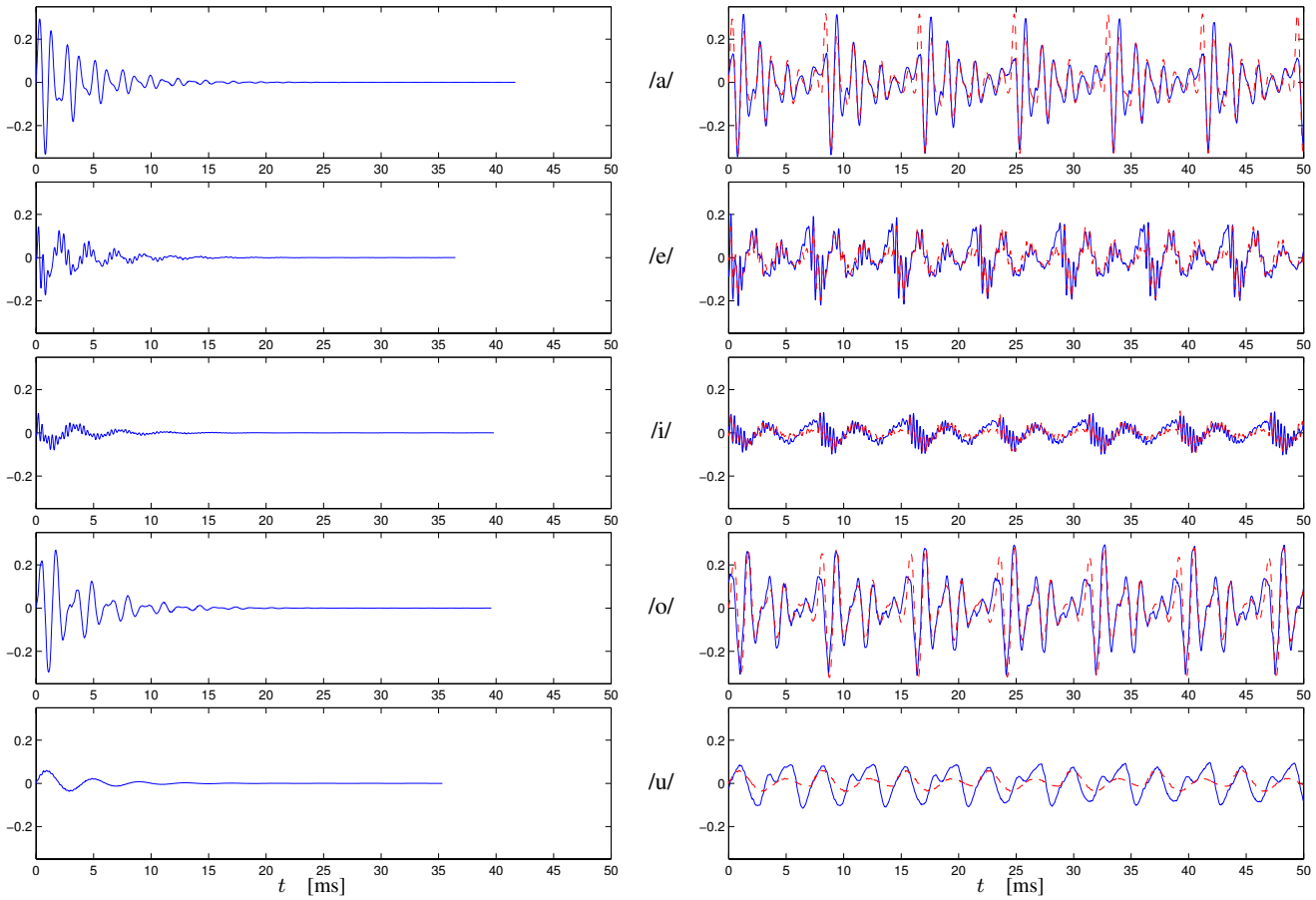


Figure 5: *The five main Italian vowels: the estimated impulse response is shown in the left panel, while the synthesized (dashed line) and original (continuous line) signals are shown in the right panel.*

domain waveforms, all of the vowels produced perfectly intelligible synthetic speech signals.

7. Conclusions

In this paper a novel model for voiced speech signals, based on the modeling of the vocal-tract impulse response by damped Bessel functions derived from an orthogonal transform pair, is presented. A modeling technique that exploits interesting properties of the employed transform is also discussed, and experimental results demonstrating its effectiveness to model Italian vowels are shown.

The proposed model seems suitable for very-low-bit-rate speech coding applications, and preliminary informal listening tests suggested that a quality superior to that of many currently adopted speech vocoders could be achieved at a comparable or lower bit rate. The extension of the technique to unvoiced sounds is currently being investigated.

8. References

[1] J. D. Gibson, "Speech coding methods, standards, and applications," *IEEE Circuits Syst. Mag.*, vol. 5, no. 4, pp. 30–49, 2005.

[2] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[3] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 3, pp. 121–132, Mar. 2004.

[4] R. Boyer and K. Abed-Meraim, "Damped and delayed sinusoidal model for transient signals," *IEEE Trans. Signal Processing*, vol. 53, no. 5, pp. 1720–1730, May 2005.

[5] C. Chen, K. Gopalan, and P. Mitra, "Speech signal analysis and synthesis via Fourier-Bessel representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '85)*, Apr. 1985, vol. 10, pp. 497–500.

[6] K. Gopalan, "Speech coding using Fourier-Bessel expansion of speech signals," in *Proc. 27th Annual Conf. IEEE Industrial Electronics Society (IECON '01)*, Nov. 2001, vol. 3, pp. 2199–2203.

[7] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: A history of the cepstrum," *IEEE Signal Processing Mag.*, vol. 21, no. 5, pp. 95–106, Sept. 2004.