The Effect of Spectral Estimator on Common Spectral Measures for Sibilant Fricatives

Patrick Reidy[1] *and Mary Beckman*[1]

[1]Department of Linguistics, Ohio State University, Columbus, OH, USA

reidy@ling.ohio-state.edu, mbeckman@ling.ohio-state.edu

Abstract

Recently, speech researchers have begun to base spectral analyses of sibilant fricatives on modern spectral estimators that promise reduced error in the estimation of the spectrum of the acoustic waveform. In this paper we look at the effect that the choice of spectral estimator has on the estimation of spectral properties of English voiceless sibilant fricatives.

Index Terms: spectral estimation, multitaper spectrum, sibilants

1. Introduction

From a statistical point of view the spectral analysis of speech data necessarily involves the estimation of the spectral density of a stochastic process that is realized as the acoustic waveform [1]. The statistical literature offers many spectral estimation methods based on the discrete Fourier transform (DFT). The spectral estimator that is most familiar to speech researchers is the windowed periodogram, which is equal to a scaled DFT.

Definition 1.1 (Windowed periodogram)**.** If $x_0, \ldots, x_{N-1}$ are data realized from a stationary process and $h_0, \ldots, h_{N-1}$ is a data window, then the *windowed periodogram* I_x of the data is defined by

$$I_x(\omega_j) = N^{-1} \sum_{n=0}^{N-1} h_n x_n e^{-2\pi i \omega_j n}, \qquad (1)$$

where $\omega_j = j/N$ for $j = 0, \ldots, N-1$. In the case where $\{h_n\}$ is a Hamming window, we refer to (1) as the Hamming-windowed periodogram.

Despite its prevalence in speech research, the windowed periodogram is known to be a poor estimator of the spectral density. More specifically, each ordinate $I_x(\omega_j)$ of the periodogram is known to be a poor point estimate of the value of the spectral density at frequency ω_j, each $I_x(\omega_j)$ susceptible to large variance and bias, resulting in a large mean square error [2].

An alternative spectral estimator to the windowed periodogram is the multitaper spectrum [3], which shares an affinity with Welch's time-averaging method: divide the data into K subintervals; compute the windowed periodogram of each subinterval; and then average these K periodograms pointwise, yielding a spectral estimate, each of whose ordinates has $1/K^{th}$ the variance of the corresponding ordinate of the ordinary periodogram [4].

The computation of the multitaper spectrum differs from the time-averaging method in two important respects: First, K copies of *all* the data are used. Second, each copy of the data is modified by a *different* windowing function, drawn from a special class of signals called the discrete prolate spheroidal sequences (DPSS) [5], [6]. (See Chapter 8 of [2] for how to compute the values of a DPSS window.)

The DPSS windows have two properties that are crucial to their use in the computation of the multitaper spectrum. The first regards the distribution of a DPSS window's energy across frequencies. If $x_0, \ldots, x_{N-1}$ is a finite sequence whose spectrum is X, and W is a real number such that $0 < W$, then the spectral concentration of X in the frequency band $[-W, W]$, denoted by $\lambda(N, W)$, is

$$\lambda(N, W) = \frac{\int_{-W}^{W} |X(\omega)|^2 \, \mathrm{d}\omega}{\int_{-\infty}^{\infty} |X(\omega)|^2 \, \mathrm{d}\omega}. \qquad (2)$$

It turns out that the finite sequences of length N can be ranked according to their concentration in $[-W, W]$. We use this fact to define the DPSS windows.

Definition 1.2 (DPSS window)**.** Given a fixed number N of elements in a sequence and a fixed frequency W, the *DPSS window of order* m, denoted by $v_0^{(m)}, \ldots, v_{N-1}^{(m)}$, is the sequence of length N with the $(m+1)^{th}$ maximal concentration $\lambda(N, W)$ in the frequency band $[-W, W]$.

The second property of the DPSS windows that is germane to spectral estimation is that DPSS windows of different order are orthogonal in the sense that

$$\sum_{n=0}^{N-1} v_n^{(j)} \cdot v_n^{(k)} = 0, \qquad (3)$$

when $j \neq k$. This ensures that the DPSS-windowed periodograms averaged to compute the multitaper spectrum

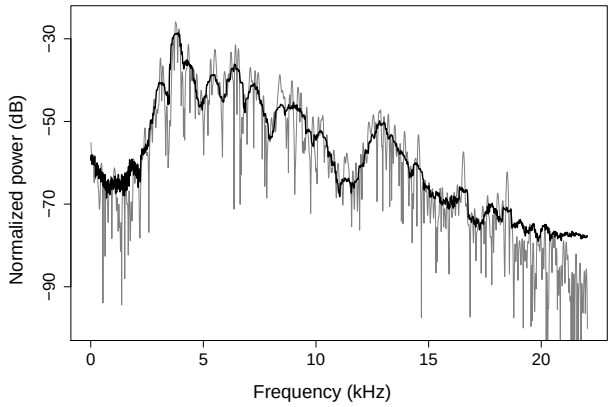

Figure 1: A Hamming-windowed periodogram (gray) and an eighth-order multitaper spectrum (black), each computed from the center 20 ms of the [ʃ] in an adult female native English speaker's production of 'ship'.

are mutually uncorrelated, which is necessary to establish convergence properties of the multitaper spectrum's ordinates [2], [3].

Definition 1.3 (Multitaper spectrum)**.** If $x_0, \ldots, x_{N-1}$ are data realized from a stationary process, then the *multitaper spectrum of order* K, denoted $M^{(K)}$, is defined by

$$M^{(K)}(\omega_j) = \frac{1}{K} \sum_{k=0}^{K-1} \hat{S}_k(\omega_j), \tag{4}$$

where ω_j is as it was in (1), and

$$\hat{S}_k(\omega_j) = N^{-1} \sum_{n=0}^{N-1} v_n^{(k)} x_n e^{-2\pi i \omega_j n}. \tag{5}$$

It has been shown that the ordinates of the multitaper spectrum have lower bias and variance than those of the periodogram [3]. These differences are illustrated in Figure 1, where the variance properties of each spectral estimator are revealed through the point-to-point fluctuations of its ordinate values, these fluctuations being of much lesser magnitude for the multitaper spectrum. The difference between the bias properties of each estimator is most clearly seen at high frequencies where the general trendlines of the two estimates begin to diverge. As a consequence of its ordinates having lower bias and variance compared to the periodogram, the multitaper spectrum provides a closer estimate to the spectral density, in the sense that the summed mean square error of its ordinates is less than that of the periodogram.

While there is no doubt over the multitaper spectrum's superiority over the windowed periodogram as a spectral estimator, its utility to spectral analyses of speech data is still unclear since there are many situations in which the researcher discards much of the information from the estimated spectrum, instead choosing a handful of quantities derived from the spectral estimate itself. For example, vowels are commonly described by their formants and fricatives by their spectral moments, a rough measure of their spectral shape [7], [8], [9]. Hence, for many types of speech analysis, the choice of spectral estimator will affect the final analysis only if these spectral properties differ significantly from one estimate to the next. Indeed, the comparison of the periodogram and multitaper spectrum shown in Figure 1 suggests that the choice of spectral estimator may have no significant impact on certain properties of the estimate as both estimates agree on the locations of the spectral peaks and troughs. In the remainder of the paper, we investigate whether the choice of spectral estimator—specifically, the Hamming window periodogram or the multitaper spectrum—has a significant effect on the estimation of spectral measurements that have proven to be successful in describing the contrast between the English voiceless sibilants /s/ and /ʃ/.

2. Spectral properties of sibilants

In this section, we review two spectral measures that have been shown to provide robust separation of English /s/ and /ʃ/. The articulation of both English sibilants /s/ and /ʃ/ involves a narrow constriction in the vocal tract through which turbulent air flows, creating a noise source that excites the oral cavity anterior to the constriction. Previous studies of the articulation of /s/ and /ʃ/ have found that English speakers consistently articulate /ʃ/ with a significantly larger front cavity than that of /s/ [10], [11].

This articulatory difference engenders spectral differences between /s/ and /ʃ/. The larger front cavity involved in the production of /ʃ/ causes the main resonance of the vocal tract to occur at a lower frequency than that of /s/. This relative difference in resonant frequencies is revealed spectrally as a difference in the position of the *main spectral peak*. Acoustic studies of /s/ and /ʃ/ have shown that the location of the most prominant spectral peak robustly differentiates these two sibilants [8].

Another commonly used spectral measure that has proved successful in differentiating /s/ and /ʃ/ is the *centroid*, or *spectral mean* [7]. The centroid of a spectral estimate is found by first normalizing the ordinate values of the estimate so that they sum to one. The centroid is then a weighted sum of the Fourier frequencies at which the estimate is defined, each frequency weighted by the estimate's ordinate value at that frequency.

3. Experiment

3.1. Data collection and preparation

The speakers in the current study were 18 (10 females, 8 males) native English-speaking adults. Their produc-

tions of the target sibilant fricatives, English /s/ and /ʃ/, were recorded at a sampling rate of 44.1 kHz, during a picture-prompted word repetition task. Each target fricative occurred in the initial, pre-vocalic position of 30 real words chosen so that each fricative occurred in a variety of vowel contexts. Half of these data have been analyzed in [12], and the reader is referred there for a more complete description of the recording procedure along with a full list of the target words.

English-speaking phoneticians marked the onset and fricative-vowel boundary of all but 32 of the recorded tokens, which left 508 tokens that were included in the current study. The frication onset was marked at the earliest point at which an increase in the waveform's amplitude coincided with the presence of high-frequency energy in the spectrogram; the fricative-vowel boundary, at the zero-crossing of the first upswing of the periodic portion of the waveform.

3.2. Spectral measurements

For the spectral measurements made from each token, a 883 point (≈ 20.02 ms) analysis window was centered at the token's temporal midpoint, determined from the frication onset and fricative-vowel boundary times marked during the data preparation stage. The nonstationary nature of a word-initial sibilant's acoustic waveform imposes such a short duration on the analysis window in order for the spectrum to be estimated from a stationary portion of the waveform. While the waveform of a word-initial sibilant cannot be assumed *a priori* to be more stationary at its temporal center than at its boundaries, the choice to estimate the spectrum from the center 20 ms of the waveform came from evidence suggesting that the articulators involved in the production of /s/ reach their target position approximately at the midpoint of the waveform's duration [13]. Hence, we take the center portion of the waveform as representative of a sibilant in the sense that it is the acoustic consequence of certain articulators reaching their positional target during the articulation of the sibilant.

For each token, the 883 point analysis window was extracted using a rectangular window and then pre-emphasized according to the first-order difference equation $y_n = x_n - 0.98x_{n-1}$, where x and y are the waveform before and after pre-emphasis, respectively. After pre-emphasis, two spectral estimates were computed: a Hamming-windowed periodogram and an eighth-order multitaper spectrum. The multitaper spectrum was computed with DPSS windows generated using the parameters $N = 883$ and $W = 4/883$. Each spectral estimate was high-pass filtered to remove all spectral components below 320 Hz, which would have come from ambient noise. Finally, each spectral estimate was used to estimate the peak and centroid frequencies of the sibilant's spectrum.

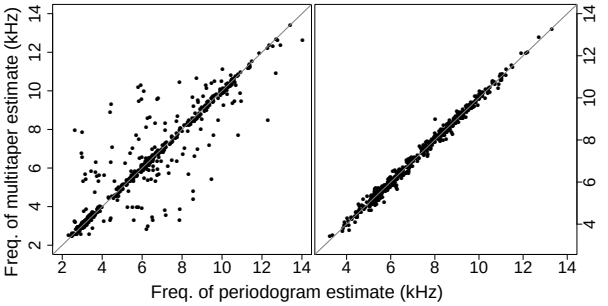

Figure 2: Comparisons of peak (left panel) and centroid (right panel) estimates. The $y = x$ line is shown in gray in each panel.

3.3. Results

The left panel of Figure 2 compares the peak frequency estimates derived from the Hamming-windowed periodogram to those computed from the multitaper spectrum. Here, the peak frequency of the multitaper spectrum estimated from a sibilant waveform is plotted against the peak frequency of the periodogram estimated from that same waveform. A strong linear correlation holds between these two methods of estimating the peak frequency, as many of the points fall on or very near to the $y = x$ line.

In order to determine whether the choice of spectral estimator affects the estimated value of a sibilant fricative's peak frequency, the peak frequency estimates computed from the periodogram were compared to those computed from the multitaper spectrum using a paired t-test. The peak frequency estimate derived from the periodogram ($\mu = 6456.537, \sigma = 2612.724$) is on average slightly lower than the same estimate computed from the multitaper spectrum ($\mu = 6480.488, \sigma = 2590.109$); however, the paired t-test revealed that this difference is not significant, $t(507) = -0.500, p \approx 0.617$. A 95% confidence interval for the difference between the periodogram's peak frequency and that of the multitaper spectrum ranges from -117.999 Hz to 70.097 Hz.

A comparison of the two different centroid estimates is shown in the right panel of Figure 2, where the centroid of the multitaper spectrum computed from a sibilant token is plotted against the periodogram's centroid, computed from the same token. The centroid estimates show even less deviation from the $y = x$ line than did the peak frequency estimates.

The centroid estimates computed from each spectral estimator were subjected to a paired t-test in order to determine whether the estimated value of a sibilant fricative's spectral mean is significantly affected by the choice of spectral estimator. This paired t-test revealed that the difference between the periodogram's centroid ($\mu = 7263.034, \sigma = 1908.658$) and the multitaper spectrum's centroid ($\mu = 7259.717, \sigma = 1918.366$) is not

statistically significant, $t(507) = 0.4267, p \approx 0.6698$. A 95% confidence interval for the difference between the centroid of the periodogram and that of the multitaper spectrum is very narrow, ranging only from -11.956 Hz to 18.591 Hz.

3.4. Discussion

Some researchers have recently begun to use the multitaper spectrum in their analyses of sibilant fricatives (e.g., [13], [14]), some implying that the periodogram is too errorful of an estimate for a valid analysis of these sounds [15], [16]. If this implication held true, then comparisons between these new studies and older studies that estimated the spectrum via the periodogram would be impossible. Indeed, the results of these older studies would seem to be invalid.

Our results suggest that for spectral measures that are commonly used in alayses of sibilant fricatives, the type of spectral estimate should not be expected to significantly influence the outcome of the analysis. Jongman, Wayland, and Wong found the spectral peak of alveolar sibilants (6839 Hz for /s/ and /z/ together) to be approximately 3000 Hz higher than that of post-alveolars (3820 for /ʃ/ and /ʒ/ together); and the centroid of alveolar sibilants (6133 Hz) to be approximately 2000 Hz higher than that of post-alveolars (4229 Hz) [8]. We found that when estimated with the multitaper spectrum, the spectral peak and the centroid frequency are individually at least 95% likely to be less than 120 Hz and 20 Hz, respectively, of their periodogram-estimated values.

While our results suggest that a switch to using the multitaper spectrum is unlikely to affect the estimates of spectral properties of sibilant fricatives, we do not intend for this to be taken as an implication that spectral estimates other than the periodogram have no place in an analysis of sibilant fricatives. In fact, we think that there is good reason to use the multitaper spectrum instead of the periodogram for visualizing sibilant spectra since the reduced variance of the multitaper spectrum's ordinates makes it easier to distinguish visually trends in the contour of the spectrum. For example, the multitaper spectrum in Figure 1 shows three small, but clearly defined peaks in the $\approx$ 5000–8000 Hz region, just above the peak frequency. Looking at this same frequency range in the periodogram, however, the large variance of its ordinates makes it unclear whether these are true peaks or whether they are insignificant fluctuations from the downward sloping trendline of the spectrum.

4. Conclusion

In this paper we investigated the effect of spectral estimator on the estimate of spectral peak and centroid frequencies, two properties that have been shown to robustly differentiate the English sibilant fricatives /s/ and /ʃ/. We found no significant difference between the estimates computed from a multitaper spectrum and those derived from a periodogram, suggesting that analyses that use either spectral estimator to compute these spectral measurements are comparable and equally valid.

5. Acknowledgements

Data collection was supported by NIDCD grant R01 02932 to Jan Edwards, and this work was supported by funding from the OSU Center for Cognitive Science. Thanks to Fangfang Li and Chanelle Mays for help with data preparation.

6. References

[1] Shumway, R. H. and Stoffer, D. S., *Time Series Analysis and Its Applications*. Springer, 2nd edition, 2006.

[2] Percival, D. B. and Walden, A. T., *Spectral Analysis for Physical Applications*. Cambridge Univeristy Press, 1993.

[3] Thomson, D. J., "Spectrum estimation and harmonic analysis", *Proc. of the IEEE*, 70:1055–1096, 1982.

[4] Welch, P. D., "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms", *IEEE Trans. Audio and Electroacoustics*, 15(2):70–73, 1967.

[5] Slepian, D. and Pollack, H. O., "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—I", *Bell System Tech. Journal*, 40:43–64, 1964.

[6] Slepian, D., "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty—V: The Discrete Case", *Bell System Tech. Journal*, 57(5):1371–1430, 1978.

[7] Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R. N., "Statistical analysis of word-initial voiceless obstruents: Preliminary data", *J. Acoust. Soc. Am.*, 84:115–124, 1988.

[8] Jongman, A., Wayland, R., and Wong, S., "Acoustic characteristics of English fricatives", *J. Acoust. Soc. Am.*, 108(3):1252–1263, 2000.

[9] Li, F., Munson, B., Edwards, J., Yoneyama, K., and Hall, K., "Language specificity in the perception of voiceless sibilant fricatives in Japanese and English: Implications for cross-language differences in speech-sound development", *J. Acoust. Soc. Am.*, 129(2):999—1011, 2011.

[10] Narayanan, S. S., Alwan, A. A., and Haker, K., "An articulatory study of fricative consonants using magnetic resonance imaging", *J. Acoust. Soc. Am.*, 93(3):1325–1347, 1995.

[11] Toda, M. and Honda, K., "An MRI-based cross-linguistic study of sibilant fricatives", *Proc. 6th ISSP*, 2003.

[12] Li, F., "Language-Specific Developmental Differences in Speech Production: A Cross-Language Acoustic Study", *Child Dev.*, In press.

[13] Iskarous, K., Shadle, C. H., and Proctor, M. I., "Articulatory-acoustic kinematics: The production of American English /s/", *J. Acoust. Soc. Am.*, 129(2):944–954, 2011.

[14] Todd, A. E., Edwards, J. R., Litovsky, R. Y., "Production of contrast between sibilant fricatives by children with cochlear implants", *J. Acoust. Soc. Am.*, 130(6):3969–3979, 2011.

[15] Blacklock, O. S., "Characteristics of viariation in production of normal and disorded fricatives, using reduced-variance spectral methods", Ph. D. dissertation, University of Southampton, 2004.

[16] Shadle, C. H., "Phonetics, Acoustic", *Encyclopedia of Lang. & Ling.*, ed. by K. Brown. Elsevier, 2nd ed., vol. 9, pp. 442–460, 2006.