Spectral Moment vs. Bark Cepstral Analysis of Children's Word-initial Voiceless Stops

H. Timothy Bunnell, James Polikoff, and Jane McNicholas

Speech Research Laboratory
Alfred I. duPont Hospital for Children, Nemours Children's Clinic
Wilmington, Delaware, USA
bunnell@asel.udel.edu

Abstract

Spectral moments analysis has been shown to be effective in deriving acoustic features for classifying voiceless stop release bursts [1], and is an analysis method that has commonly been cited in the clinical phonetics literature dealing with children's disordered speech. In this study, we compared the classification of stops /p/, /t/, and /k/ based on spectral moments with classification based on an equal number of Bark Cepstrum coefficients. Utterance-initial /p/, /t/, and /k/ (1338 samples in all) were collected from a database of children's speech. Linear discriminant analysis (LDA) was used to classify the three stops based on four analysis frames from the initial 40 msec of each token. The best classification based on spectral moments used all four spectral moment features (computed from bark-scaled spectra) and all four time intervals and yielded 75.6% correct classification. The best classification based on Bark cepstrum yielded 83.4% correct also using four coefficients and four time frames.

1. Introduction

Spectral moments analysis has become a popular method of analysis for obstruent segments, especially in the literature on clinical phonetics [1-5]. The argument presented originally in [1] for preferring spectral moments features over other features of spectral shape was that the relative and non-dimensional nature of spectral moments allowed them to better capture the important spectral shape features in a manner that varied less across talkers and individual utterances. Indeed, the results reported in [1] suggested that spectral moments afforded more accurate classification of instances of stop bursts than did acoustic feature sets derived from LPC analyses. Subsequent work suggests that spectral moments do not provide an adequate characterization of vowels and continuents [6], but for obstruent and stop spectra spectral moments appear to capture adequate information to support segment classification and to distinguish subtle differences between normal and distorted segment productions [2].

Another cited advantage of using spectral moment-based acoustic features is that their computation is straightforward and unambiguous. This is a characteristic shared by other acoustic analysis techniques, notably the cepstral analysis techniques commonly used in the speech recognition literature [7]. Perceptually weighted (Mel or Bark scaled) cepstral features are the basis for acoustic feature sets in a majority of recent ASR publications and clearly provide useful characterizations of speech acoustics for vowels and continuents as well as obstruents. However, there do not appear to be any reports directly comparing spectral moments with cepstral features for classification of obstruents. It is possible that spectral moments may perform better than cepstral features for classifying some classes of segments. If so, this could lead to improvements in feature extraction for ASR and would also serve to validate the use of spectral moments analyses in the clinical phonetics literature. On the other hand, should cepstral features prove equally or more effective for classifying obstruent spectra, it may lead to recommendations to alter analysis techniques common among clinical phoneticians and speech language pathologists.

To examine this issue, the present study directly compared cepstral features and spectral moments features for the classification of burst spectra from utterance-initial voiceless plosives /p/, /t/, and /k/. An additional factor we considered in the present study is the observation that virtually all the published reports involving the use of spectral moments analyses have been based on a relatively small number of individual talkers and speech tokens. Thus, published accounts of classification accuracy for spectral moment features may overestimate the accuracy that would be observed for a larger sample of talkers and tokens. To address this concern, we analyzed tokens from a group of around 200 children aged 6 to 8 whose recordings had been sampled for a children's speech database.

2. Method

2.1. Subjects

The subjects involved in this study were a group of 208 children, whose ages ranged from six to eight years old. Each subject recorded a series of 100 individual English words in isolation for a corpus of children's speech that was recorded as part of an unrelated project in the Speech Research Laboratory at the A. I. duPont Hospital for Children in Wilmington, Delaware.

2.2. Stimuli

Burst segments were extracted for the voiceless stop consonants /p/, /t/, and /k/ from the corpus of speech obtained from the subjects. Bursts were only extracted from voiceless stop consonants occurring in word-initial position and an attempt was made to balance phonemic context such that for each class of voiceless stop, the number and type of following phonemes occurred in roughly equal numbers. This resulted in a balanced set with 446 bursts to be analyzed for each stop. Each burst that was extracted was aligned so that the burst started at 20msec from the beginning of the waveform file. This was accomplished automatically by a program that

recognized the burst in the original waveform file (containing the recording of the full word) and copied only the burst to a second waveform file, padding silence to the beginning and end of the file to ensure that the burst began 20msec from the file onset and that the total file was 100msec long. After the program extracted the burst segments, each file was examined manually to verify that it contained a correctly aligned /p/, /t/, or /k/ burst segment. Any alignment errors detected in this process were hand-corrected.

2.3. Procedure

The moments program [8] was used to analyze each burst. Spectra windows were adjusted so that the burst onset was located at 20ms and the entire window was set at 100ms. The output variables of the moments program were mean, skewness, kurtosis, and variance. For each burst in the dataset, spectral moments were calculated with a 20ms window centered on the analysis point at four different time intervals (20, 30, 40, and 50ms). These were subjected to a linear discriminant analysis (LDA) with the stop consonant (/p/, /t/, or /k/) as the grouping variable. A second LDA was performed using moments computed from Bark-scaled spectra. Both of these analyses were run with and without the use of variance as a variable in the analysis. Analyses reported in [1] and subsequent reports often omit use of the variance component as not making a significant independent contribution to classification.

A further analysis of the dataset was performed using a Bark Cepstrum analysis program developed locally. In this analysis, six cepstral coefficients (DC and first five cosine terms) were estimated for each frame. The analysis program (available at http://www.asel.udel.edu/speech) computes log energy in each of 32 bands evenly spaced on a Bark scale. Each band is triangular in shape and overlaps adjacent bands by 50%. These analyses used the same 20 ms window and 10 msec stepping parameters that were used for moment analyses. Features derived from these analyses were also analyzed using LDA with the stop consonant (/p/, /t/, or /k/) as the grouping variable.

3. Results

For /p/, the percentage of phonemes correctly classified using linear moments and input variables of mean, skewness, kurtosis, and variance was 62.3, 71.1, 72.6, and 72.9% for the first, first plus second, first, second and third, and all four time frames of the burst respectively. When Bark-scaled moments were used in the analysis, the percent correct classification at these intervals was 73.3, 83.9, 82.1, and 80.5%, with classification accuracy dropping somewhat when the fourth frame was added into the analysis. When variance was excluded as a variable in the analysis, the percent correct for linear moments was 57.6, 67.3, 70.2, and 71.1% and Bark-scaled moments yielded 78.0, 78.9, 76.9, and 77.4 percent correct. Thus, for /p/, the best classification accuracy was obtained with variance included and using only the first three of the four analysis frames.

The percentage of the phoneme, /t/, correctly classified using linear moments and input variables of mean, skewness, kurtosis, and variance was 69.5, 87.0, 86.3, and 86.5% for the first, first+second, first+second+third, and all four frames of the burst respectively. When Bark-scaled moments were used in the analysis, the percent correct classification at these intervals was 69.1, 81.8, 82.1, and 80.7%. When variance was excluded as a variable in the analysis, the percent correct for linear moments was 69.5, 86.5, 86.1, and 85.9% and Bark-scaled moments yielded 70.2, 84.5, 85.0, and 84.5 percent correct. Thus, for /t/, the best classification was obtained for the LDA using all four moments (linear scale), but only two analysis frames, the first and second.

For the phoneme /k/, the percent correctly classified using linear moments and input variables of mean, skewness, kurtosis, and variance was 57.6, 61.2, 64.1, and 64.6% for the first 20, 30, 40, and 50ms of the burst respectively. When Bark-scaled moments were used in the analysis, the percent correct classification at these intervals was 57.2, 60.8, 61.4, and 65.7%. When variance was excluded as a variable in the analysis, the percent correct for linear moments was 58.5, 64.1, 61.7, and 62.8% and Bark-scaled moments yielded 51.1, 54.3, 54.7, and 61.7 percent correct. For /k/, best performance was obtained with all four (linear) moments and all four time intervals combined.

When collapsed across target phoneme, the overall percentage of voiceless stops correctly classified using linear moments and input variables of mean, skewness, kurtosis, and variance was 63.2, 73.1, 74.4, and 74.7% for the first 20, 30, 40, and 50ms of the burst respectively. When Bark-scaled moments were used in the analysis, the overall percent correct classification at these intervals was 66.5, 75.5, 75.2, and 75.6%. When variance was excluded as a variable in the analysis, the overall percent correct for linear moments was 61.9, 72.6, 72.6, and 73.2%, while Bark-scaled moments yielded 66.4, 72.6, 72.2, and 74.5 overall percent correct. Thus, overall, the best observed classification (75.6% correct) was obtained using all four bark-scaled moments at all four analysis frames.

For equivalence with the 4-term moment analyses, four cepstral parameters were selected (the zeroeth or DC component, plus the second, fourth and fifth coefficients). Combined over the four time frames, these four coefficients supported LDA classification of the stop bursts with an overall accuracy of 83.0 percent correct. We also calculated LDA classifications using the first six Bark-cepstrum coefficients. The percentage of /t/ phonemes correctly classified was 63.2, 81.6, 82.3, and 82.3% respectively when data from analysis windows at 20, 30, 40, and 50ms were added to the analysis. For /k/, the corresponding percent correct was 63.7, 75.8, 77.6, and 78.7%. And for /p/ the percent correct was 74.0, 87.9, 87.9, and 89.2% as data from each successive window was added to the analysis. When collapsed across target phoneme, the overall percentage of voiceless stops correctly classified using Bark-cepstrum analysis was 67.0, 81.8, 82.6, and 83.4% as each window was added to the analysis.

4. Discussion

As with previous analyses of stop release bursts (e.g., [1, 9]), we found that information in successive analysis frames distributed over the release burst contributes independently to accurate classification of stops. Using multiple analysis frames leads to better classification than is available from any single analysis frame. Unlike the initial reports of spectral moment analyses [1] which indicated that variance did not contribute to classification accuracy, we found generally better classification accuracy when all four moments were

used. Our results also differed from the original report in finding that the Bark features lead to slightly better overall performance than did linear frequency based moments. It is most likely that the much larger number of individual talkers and overall number of tokens is responsible for the difference we observed in whether variance contributed significantly to classification accuracy, if only because the much larger N provides much greater power to detect small differences. The larger N cannot account for the finding in the present analysis that Bark rather than linear frequency scales lead to better performance.

Perhaps the most important result of the present analyses, however, is the finding that Bark Cepstral features perform better than do spectral moments in overall classification accuracy. Our laboratory is presently developing normative monophone HMMs for segments uttered by young children. These models, which are based on our Bark Cepstrum features, are used to assess progress in speech training for children with speech disorders. Based on the results of the present study, we suspect that the Bark Cepstrum features will afford better classification of children's speech accuracy that would be available using spectral moments despite the prevalence of that analysis approach in much of the clinical speech literature.

5. Acknowledgements

The recording process for the database of children's speech used in this study was supported by Voiceware Co., Ltd. of Seoul, Korea. The authors would like to thank Jenna Hammond for her assistance with preliminary data analysis. The authors also wish to thank Rachel Maslow and Susan Ramsey for their help with data collection. This work was supported by funding from the Nemours Foundation.

6. References

[1] K. Forrest, G. Weismer, P. Milenkovic, and R. N. Dougall, "Statistical analysis of word-initial voiceless obstruents: preliminary data," *J Acoust Soc Am*, vol. 84, pp. 115-23., 1988.

[2] K. Forrest, G. Weismer, M. Hodge, D. A. Dinnsen, and M. Elbert, "Statistical-Analysis of Word-Initial K and T Produced by Normal and Phonologically Disordered Children," *Clinical Linguistics & Phonetics*, vol. 4, pp. 327-340, 1990.

[3] K. Forrest, G. Weismer, M. Elbert, and D. A. Dinnsen, "Spectral-Analysis of Target-Appropriate T and K Produced by Phonologically Disordered and Normally Articulating Children," *Clinical Linguistics & Phonetics*, vol. 8, pp. 267-281, 1994.

[4] P. Flipsen, Jr., L. Shriberg, G. Weismer, H. Karlsson, and J. McSweeny, "Acoustic characteristics of /s/ in adolescents," *J Speech Lang Hear Res*, vol. 42, pp. 663-77., 1999.

[5] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives," *Journal of the Acoustical Society of America*, vol. 108, pp. 1252-1263, 2000.

[6] P. Flipsen, Jr., L. D. Shriberg, G. Weismer, H. Karlsson, and J. McSweeny, "Acoustic phenotypes for speech-genetics studies: reference data for residual /Er/ distortion," *Clin linguist Phon*, vol. 15, pp. 603-630, 2001.

[7] J. Deller, R., J. Proakis, G. , and J. Hanson, H., L., *Discrete-Time Processing of Speech Signals*: MacMillan, 1993.

[8] P. Milenkovic, "Moments: batch speech spectrum moments analysis." Madison, Wisconsin, 1999.

[9] D. Kewley-Port, "Time-Varying Features as Correlates of Place of Articulation in Stop Consonants," *Journal of the Acoustical Society of America*, vol. 73, pp. 322-335, 1983.