Spectral Moments Versus Bark Cepstrum Classification of Children's Voiceless Stops

James Polikoff, Jenna Hammond, Jane McNicholas, and H. Timothy Bunnell
Speech Research Laboratory, Alfred I. duPont Hospital for Children, Nemours Children's Clinic, Wilmington, DE 19803

ABSTRACT

Spectral moments have been shown to be effective in deriving acoustic features for classifying voiceless stop release bursts [K. Forrest, G. Weismer, P. Milenkovic, & R.N. Dougall, J. Acoust. Soc. Am, 84(1), 115-123 (1988)]. 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. The speech tokens were 446 instances each of utterance-initial /p/, /t/, and /k/ sampled from utterances produced by 208 children 6 to 8 years old. 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 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. Differences between these results and previous classification results using spectral moments will be discussed. Implications for future research on the acoustic characteristics of children's speech will be considered.

INTRODUCTION

Spectral moments analysis has become a popular method of analysis for obstruent segments, especially in the literature on clinical phonetics [1-5]. This type of analysis treats the acoustic speech spectrum as a distribution of energy, the shape of which can be characterized in terms of its mean, variance, skewness and kurtosis. The argument presented originally in [1] for preferring spectral moments features over other features of spectral shape was that their relative and non-dimensional nature allowed them to better capture the important gross features of spectral shape with less variability 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 when compared with feature sets like formant frequencies, which cannot always be unambiguously identified in the acoustic speech spectrum. However, computational ease and non-ambiguity are characteristics shared by other acoustic analysis techniques, notably the Cepstral analysis techniques commonly used to derive feature sets for speech recognition [7]. Furthermore, perceptually weighted (Mel- or Bark-scaled) Cepstral features clearly provide useful characterizations of speech acoustics for vowels and continuents as well as obstruents.

To date, there do not appear to be any reports directly comparing spectral moments with perceptually weighted 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.

METHODS

Subjects. 208 children, whose ages ranged from six to eight years recorded a series of 100 individual English words in isolation. All children were reported by parents to be developing normally and with no history of hearing or speech disorders.

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 20 msec 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 20 msec from the file onset and that the total file was 100 msec 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. As an example, Figure 1 shows a burst extracted from the initial [k] of the word *cultivate* spoken by a 7-year-old girl. The spectral cross section in this figure is the LPC smoothed spectrum obtained from a 20 msec analysis window centered on the release burst.

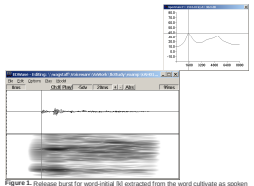

Figure 1. Release burst for word-initial [k] extracted from the word cultivate as spoken by a 7-year-old girl. The vertical line in the waveform spectrogram display indicates the location from which the spectral cross section (smaller panel) was computed using an LPC analysis with 20 msec window.

Procedure. Two acoustic analysis techniques were applied to the burst data. First, the moments program [8] was used to derive measures of the mean, skewness, kurtosis, and variance in a sequence of four frames based on 20 msec windows beginning with a frame centered on the burst release and stepping through the subsequent friction and aspiration in 10 msec steps. The moments program computes measures for both linear frequency and Bark frequency spectral representations.
The second acoustic analysis duplicated the framing parameters of the moments analysis using a Bark Cepstrum analysis program developed locally. In this analysis, six Cepstral coefficients (DC and first five cosine terms – see Figure 2) were estimated for each frame. The analysis program 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%.

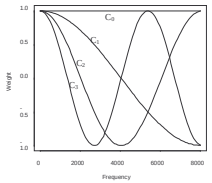

Figure 2. First four terms of Cepstrum indicating their relation to spectrum energy.

Parameters from both acoustic analyses were used in a linear discriminant analysis (LDA) with the stop consonant (/p/, /t/, or /k/) as the grouping variable. In addition to doing separate LDA analyses for the linear and Bark frequency moments, these data 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.

RESULTS

Results of the moments analysis are presented in Tables 1 and 2. Table 1 shows the results of all LDA analyses broken out by phoneme identity. Data in Table 2 are the averages collapsing over phoneme from Table 1. Thus, overall, the best observed classification (75.6% correct) was obtained using all four bark-scaled moments at all four analysis frames.

Table 1. Percentage correct classification per phoneme based on LDA analysis using parameters from 1 (Burst Only), 2 (Burst+10), 3 (Burst+10+20), or 4 (Burst+All) consecutive analysis frames. Data in each cell are p/t/k percentages.

	Burst Only	Burst+10	Burst+10+20	Burst+All
Linear with Variance	62.3/69.5/57.6	71.1/87.5/61.2	72.6/86.3/64.1	72.9/86.5/64.6
Bark with Variance	73.3/69.1/57.2	83.9/81.8/60.8	82.1/82.1/61.4	80.5/80.7/65.7
Linear w/o Variance	57.6/69.5/58.5	67.3/86.5/64.1	70.2/86.1/61.7	71.1/85.9/62.8
Bark w/o Variance	78.0/70.2/51.1	78.9/84.5/54.3	76.9/85.0/54.7	77.4/84.5/61.7

Table 2. Percentage correct classification. Data are those of Table 1, averaged over phoneme identity.

	Burst Only	Burst+10	Burst+10+20	Burst+All
Linear with Variance	62.3	73.1	74.4	74.4
Bark with Variance	66.5	75.5	75.2	75.6
Linear w/o Variance	61.9	72.6	72.6	73.2
Bark w/o Variance	66.4	72.6	72.2	74.5

Table 3 shows the results of corresponding LDA analyses using Bark Cepstral coefficients. In the first two rows of this table, percentage correct classification is shown separately for each stop (row 1) and averaged over phonemes (row 2). Since there were six Cepstral coefficients and only four spectral moments, it was possible that the generally better performance of the Bark Cepstral features was due to the larger number of degrees of freedom for the analysis. Consequently, a final LDA was run in which only four Bark Cepstral coefficients were used. The four chosen were those that carried the greatest weight in the 6-term LDA analysis (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% correct, just slightly lower than the six-term solution.

Table 3. Percentage correct consonant classification from LDA analyses using Bark Cepstral coefficients. The first row shows data broken out by phoneme (/p/t/k percentages). The second row presents the average percentage correct classification overall phonemes. The four-parameter fit is shown in the 3rd row. It covered all analysis frames and represents the percentage correct averaged over phonemes.

	Burst Only	Burst+10	Burst+10+20	Burst+All
Six parameter fit	74.0/63.2/63.7	87.9/81.6/75.8	87.9/82.3/77.6	89.2/82.3/78.7
Overall	67.0	81.8	82.6	83.4
Four parameter fit				83.0

DISCUSSION

Spectral moments analyses are gaining relatively wide use as a means of characterizing spectral shape for stop bursts and obstruent phonemes. In the present analysis, we compared the classification of word initial aspirated stop release bursts based on spectral moment features with classification based on Cepstral coefficients. The Bark Cepstral features clearly supported better classification than did the spectral moments features.

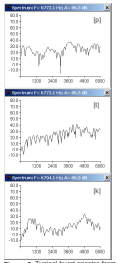

Figure 3. Typical burst spectra from the child speech database used for this study.

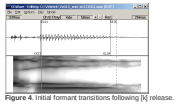

Figure 4. Initial formant transitions following [k] release.

To consider why this might be the case, we can examine examples of typical stop release burst spectra from the speech corpus we used (see Figure 3). One notable feature of spectral moments is their inability to capture finer features of spectral shape, especially when the shape is characterized as having more than one peak. As Figure 3 illustrates, [k] spectra often appeared to contain two distinct strong peaks. In this figure, the lower frequency peak is located at around 1000 Hz, with a secondary peak at around 4800 Hz. Similarly, the [k] release burst illustrated in Figure 1 shows two distinct peaks, one at around 1600 Hz, and another around 4800 Hz. Examination of spectrograms from [k] releases (Figure 4) suggests that the lower frequency peak in the [k] release may be in the neighborhood of the subsequent vowel's F2 frequency, at least for back vowels.

Another factor to consider in comparing these results with previous studies that compared spectral moments with other feature sets is the number of talkers and tokens involved. Much of the prior work with spectral moments analyses has been based on a relatively small number of tokens and speakers. The present results, because they are based on a much larger number of talkers and tokens, may better capture the relative merits of the alternative analysis methods.

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.