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 model based on spectral moments used RMS amplitude plus all four bark-scaled spectral moment features at all four time intervals and yielded 78.0% correct discrimination. The best model of similar rank based on Bark cepstrum features yielded 86.6% correct segment discrimination.

INTRODUCTION

Spectral moments analysis, which describes speech spectrum shape in terms of its mean, variance, skewness, and kurtosis, has become a popular method of analysis for obstruent segments, especially in the literature on clinical phonetics [1-5]. Moments are attractive as spectral features because they are easy and unambiguous to calculate, have been shown to provide better segment discrimination than LPC coefficients for some segments [1], and have been shown to be useful in detecting subtle yet important differences in obstruents and fricatives produced by children and adults (e.g., [2-5]).

However, there are limitations to the range of phonetic segments for which spectral moments are believed to be appropriate [6], and, since the initial report of [1] we are not aware of any that have directly compared spectral moments to other equally tractable acoustic feature sets. In particular, there has been no direct comparison of spectral moments with the Mel or Bark cepstral feature sets that are commonly used as acoustic features for speech recognition [7].

The present study directly compared cepstral features and spectral moments features for the discrimination and classification of burst spectra from utterance-initial voiceless plosives /p/, /t/, and /k/. Additionally, we sought to approach this comparison using a much larger number of speakers and tokens than previous studies have reported, and to use statistical methods that would afford a better sense of the generality of our results [8]. Thus, for the present analyses, we report both immediate "discrimination" results, that is, how well models based on the acoustic feature sets discriminated cases within the full dataset on which they were trained, and also the results of 10-fold cross-validation of the models in which results are reported for classification of unseen cases.

METHOD

Subjects

The subjects 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.

Stimuli

Burst segments were extracted from word-initial voiceless stop consonants /p/, /t/, and /k/ 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 extracted burst was aligned so that the burst started at 20msec from the beginning of the waveform file (see Figure 1 for an example). Silence was padded to the end of the file to ensure that the total file was 100msec long.

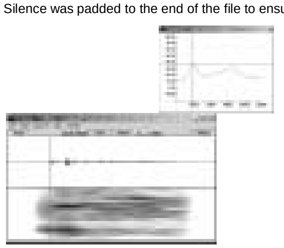

Figure 1. Release burst for word-initial [k] extracted from the word cultivate as spoken by a seven-year-old girl. The vertical line in the waveform spectrogram display indicates the location from which the spectral cross section (smaller panel) was computed using an LPC analysis with 20 msec window.

Procedure

Two acoustic analysis techniques were applied to the burst data. First, the moments program [9] was used to compute linear- and bark-frequency spectral moments 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.

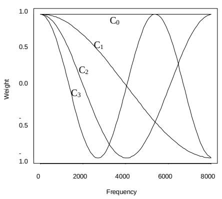

Figure 2. First four terms of Cepstrum indicating their relation to spectrum energy.

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.

Parameters from both acoustic analyses were used in a linear discriminant analysis (LDA) with the stop consonant identity (/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 obstruent classification.

RESULTS

Results are presented first for LDA discrimination. Table 1 shows the results for spectral moments data. All analyses used the RMS amplitude of the associated frame plus three or four spectral moments (i.e., a maximum of five parameters per frame). With just one exception, including the variance component in these analyses lead to better discrimination. With two exceptions, Bark-frequency moments data led to better stop discrimination than did linear-frequency moments.

Table 2 shows the results of corresponding LDA analyses using Bark cepstral coefficients. As with moments analyses, including additional analysis frames leads to improved discrimination. Overall, the six Bark cepstral coefficients provided significantly better discrimination of the stops than did the best spectral moments models (87.1 versus 78.0 percent correct). To demonstrate that this was not due simply to model rank, additional LDA analyses were run in which only the first five Bark cepstral coefficients were used. These analyses show that discrimination remains substantially better than the moments models of equal rank.

Table 1. Percentage correct LDA classification. Data are averaged over phoneme identity.

	Burst Only	Burst+ 10	Burst+ 10+20	All
Linear with Variance	63.8	74.4	75.9	76.1
Bark with Variance	65.7	75.2	77.1	78.0
Linear w/o Variance	61.0	74.0	75.4	75.2
Bark w/o Variance	66.3	73.2	73.4	76.3

Table 2. 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 five-parameter fits (dropping the 6th coefficient) are shown in the 3rd row.

	Burst Only	Burst+ 10	Burst+ 10+20	All
Six parameter fit	p 77.6	89.0	89.7	90.6
	t 63.9	82.2	83.2	85.0
	k 67.7	83.0	85.0	85.9
Overall	69.7	84.8	86.0	87.1
Five parameter fit	69.5	83.9	85.7	86.6

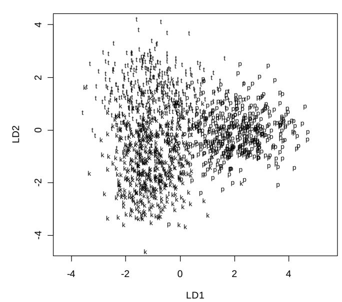

Figure 3. Position of each case on linear discriminant 1 (LD1) versus linear discriminant 2 (LD2).

Figure 3 shows the discrimination of the three stops by plotting their locations relative to the first and second linear discriminant functions for the best Bark cepstral model. The first linear discriminant primarily separates /k/ and /t/ bursts from /p/ bursts, while the second discriminant primarily separates /k/ from /t/.

Results from 10-fold cross validation of the best LDA models are shown in Tables 3 and 4. As expected, classification of unseen cases is less accurate than discrimination within the training dataset. However, the overall better performance of the Bark Cepstral feature set remains evident in these analyses.

Table 3. Predicted versus True phoneme classification for Bark-scaled spectral moments data.

	Predicted		
True	**/k/**	**/p/**	**/t/**
/k/	310	51	85
/p/	36	365	45
/t/	36	46	364

Percentage Correct = 77.65

Table 4. Predicted versus True classification for Bark-Cepstral data.

	Predicted		
True	**/k/**	**/p/**	**/t/**
/k/	376	9	61
/p/	27	400	19
/t/	44	31	371

Percentage Correct = 85.72

DISCUSSION

As with previous analyses of stop release bursts (e.g., [1, 10]), we found that information in successive analysis frames distributed over the release burst contributes independently to accurate classification of stops. 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 better overall performance than did linear frequency based moments. We attribute these differences to sampling error with the smaller dataset used by [1].

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. Given the substantial improvement in discrimination and classification performance observed here for the Bark-cepstral dataset (around 8 percent in the cross-validated analysis), we would discourage investigators from using spectral moments as acoustic features unless they wish to address specific hypotheses regarding features like the spectral mean energy or skewness. In particular, investigators interested in finding and characterizing general spectral differences, for example, to observe changes in the spectral characteristics of segments during speech training, may find that Bark-Cepstral features afford better ability to discriminate small changes than to spectral moments.

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 Nemours.

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, 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," J Acoust Soc Am, vol. 108, pp. 1252-63, 2000.

[6] P. Flipsen, L. D. Shriberg, G. Weismer, H. B. Karlsson, and J. L. McSweeny, "Acoustic phenotypes for speech-genetics studies: reference data for residual backslash 3 backslash distortions," Clinical Linguistics & Phonetics, 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] W. N. Venables and B. D. Ripley, Modern applied statistics with S, 4th ed. New York: Springer, 2002.

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

[10] D. Kewley-Port, "Time-varying features as correlates of place of articulation in stop consonants," J Acoust Soc Am, vol. 73, pp. 322-35, 1983.