INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Validation of a Task-Independent Cepstral Peak Prominence Measure with Voice Activity Detection

Olivia M. Murton[1], Abigail E. Haenssler[1,2], Marc F. Maffei[1], Kathryn P. Connaghan[1], and Jordan R. Green[1,3]

[1]Department of Communication Sciences and Disorders, MGH Institute of Health Professions, Boston, MA
[2]School of Medicine and Health Sciences, George Washington University, Washington, DC
[3]Speech and Hearing Biosciences and Technology Program, Harvard University, Cambridge, MA
oliviam@alum.mit.edu, ahaenssler@mghihp.edu, mmaffei@mghihp.edu, kconnaghan@mghihp.edu, jgreen2@mghihp.edu

Abstract

This study investigates the task-dependence of standard cepstral peak prominence (CPP) computation methods, and the advantages conferred by an open-source method of excluding unvoiced regions in CPP computation. We use Praat and a public dataset (Perceptual Voice Qualities Database, consisting of 295 speakers) to assess how well a *voice-only CPP* algorithm identifies voice disorders, identifies perceived dysphonia, and correlates with dysphonia severity. Results indicate that, compared to standard CPP computation, voice-only CPP is (1) less affected by unvoiced regions in the speech signal and (2) better reflects clinical outcomes (i.e., voice disorder diagnosis and dysphonia severity) for data sets that contain varying speech tasks. We expect voice-only CPP to be particularly useful for assessing speech that contains unknown or heterogeneous utterance types, as well as for speakers whose voice signal is affected by involvement of other speech subsystems (e.g., articulatory impairment).

Index Terms: voice, cepstral peak prominence, dysphonia

1. Introduction

Cepstral peak prominence (CPP) is an objective measure of voice quality that is widely used for both clinical and research applications. CPP has proven to be a sensitive metric for voice disorder detection that strongly correlates with perceived dysphonia severity [1]. In contrast to traditional acoustic measures of phonatory perturbation (e.g., jitter and shimmer) CPP can be calculated from connected speech and is reliable for speakers across the dysphonia severity range, including those with severe dysphonia for whom fundamental frequency cannot be directly computed [2].

Increasingly, research into clinical use of CPP has found that it is an accurate indicator of voice disorder and dysphonia severity in many contexts, including across languages, voice disorder diagnoses, speaker ages, and speaking tasks [3]–[9]. Unfortunately, the calculation of CPP is sensitive to many factors that are not directly related to voice production. In particular, choice of analysis program, specific parameter settings, and speech task can all substantially alter CPP values. Clinical measures of CPP typically divide a signal into many short time segments ("frames") and then average the CPP in each frame to produce a single mean CPP measure for an utterance. Therefore, utterances that vary in their proportion of silent or unvoiced frames (which have low CPP compared to voiced frames) can have widely varying mean CPPs even when the phonatory properties are essentially similar. As a result, normative CPP values fall in substantially different ranges for different speaking tasks. Most notably, CPP values for continuous speech are typically lower than CPP values for sustained vowels, due to the presence of silent and unvoiced frames.

Voice analysis is used in many different research and clinical contexts, often with widely varying speech stimuli. In a clinical voice evaluation, standard speech tasks may involve sustained vowels, CAPE-V sentences, or any of several reading texts [10]–[12]. Other data sets may include more naturalistic speech in the form of open-ended prompts, conversations, or entirely unprompted speech collected in daily life. As discussed above, these different speech tasks are likely to yield different CPP values depending on their phonetic content. This variety limits the generalization of findings across studies and underscores the need for more robust CPP metrics that can be used across a wide variety of speech tasks.

The sensitivity of CPP to unvoiced frames also may limit its utility for measuring dysphonia in individuals with concomitant motor speech impairments [2], which are associated with an increase in the number of silent pauses and devoicing errors [13], [14]. These errors increase the proportion of unvoiced speech and lower CPP values. Also, alterations in articulation rate unevenly affect different speech sounds, with vowels lengthening more than consonants when speech rate increases [15].

One potential solution to this problem is a CPP computation that excludes unvoiced frames. We refer to this modified CPP as the *voice-only CPP*, or vCPP. In this project, we use a publicly available dataset of speakers with and without voice disorders, accompanied by speech-language pathologists' (SLP) ratings of voice quality. We also use Praat, a freely-available speech analysis program, and a custom Praat script to compute the voice-only CPP using voice activity detection. This script is available at github.com/murtono/Praat-voice-

10.21437/Interspeech.2023-2168

only-CPP. We apply the V3 framework [16] to test the analytical and clinical validity of voice-only CPP. To that end, we address four research questions:

Analytical validity

1. Are vCPP values more robust than standard CPP across a variety of tasks?

Clinical validity

2. What is the accuracy of vCPP in identifying speakers with diagnosed voice disorders?
3. What is the accuracy of vCPP in identifying speakers with perceptually dysphonic voices?
4. What is the strength of the correlation between vCPP values and perceived dysphonia severity?

2. Methods

The Perceptual Voice Qualities Database (PVQD) consists of recordings from 295 speakers [17]. This data set is publicly available at DOI: 10.17632/9dz247gnyb.3. Of those speakers, 186 have a diagnosed voice disorder, and 89 are healthy controls, for a total of 275 participants. The remaining 20 have no diagnosis specified and are excluded from the subsequent analyses. Each recording session includes sustained vowels (/a/ and /i/; this project only used the /a/ vowel) and the six CAPE-V sentences [10]. For each session, three trained listeners provided CAPE-V [10] and GRBAS [18] ratings. Each listener rated each session twice so that reliability measures could be computed. A complete description of the methods and reliability statistics for this data set can be found in [17].

To compute standard CPP, each recording was analyzed in Praat (Version 6.2.14 retrieved at http://www.praat.org/) following the procedure described in [2]. In this method, recordings were high-pass filtered above 34 Hz using a stop Hann band, the PowerCepstrogram was computed, and the CPP (called "CPPS" in Praat) was extracted from the PowerCepstrogram. The Praat script with details for this procedure is available at github.com/murtono/Praat-voice-only-CPP. To compute voice-only CPP, recordings were again analyzed in Praat using a modification of the standard CPP procedure. First, each recording was divided into voiced and unvoiced regions using the "To TextGrid (voice activity)" Praat command. Each voiced region was extracted separately and the standard CPP was computed according to the procedure described above. The voice-only CPP was reported as the time-weighted average of the CPP of each voiced region. The Praat script used to carry out this process is included in the appendix.

Our first goal was to assess how much standard and voice-only CPP values vary with tasks of different types. Intuitively, there should be no difference between the CPP procedures when speech is completely voiced since there are no unvoiced regions to remove. Results from the two processes should diverge for stimuli with greater degrees of unvoiced time. To address this question, we compared standard and voice-only CPP for (1) sustained vowels, (2) the fully-voiced CAPE-V sentence ("We were away a year ago"), and (3) the voiceless-stop-heavy CAPE-V sentence ("Peter will keep at the peak"). The Pearson correlation coefficient between standard and voice-only CPP was computed for each of these tasks.

To address Question 2, we compared speakers with a diagnosed voice disorder against speakers with no voice disorder. Accuracy was evaluated using the area under the curve (AUC) for the receiver operating characteristic (ROC) curve. An AUC closer to 1 indicates better classification accuracy, and an AUC of 0.5 indicates chance performance. Within that range, an AUC of .7 to .8 is considered "adequate", and an AUC of .8 to .9 is considered "excellent" [19]. In this analysis, AUCs were computed for six conditions:

1. Classification of vowels based on standard CPP
2. Classification of vowels based on voice-only CPP
3. Classification of CAPE-V sentences based on standard CPP
4. Classification of CAPE-V sentences based on voice-only CPP
5. Classification of any task based on standard CPP
6. Classification of any task based on voice-only CPP

In Conditions 5 and 6, all the CAPE-V sentences and vowels were combined into a single dataset. These conditions were used to assess classification accuracy of a single CPP threshold that could be applied to any task, which is useful if the speaking task is unspecified or varies across stimuli being evaluated.

To address Question 3, we used the "Grade" section of the GRBAS perceptual ratings to group speakers into perceptually non-dysphonic ("normal" grade) or dysphonic ("mild", "moderate", or "severe" grade). Classification accuracy for this grouping was assessed using AUCs for the same six conditions described in Question 2.

To address Question 4, we used the MATLAB fitlm function to compute the linear relationship between (1) standard CPP and perceived dysphonia severity and (2) vCPP and perceived dysphonia severity. This analysis was done separately for the sustained vowel task, the CAPE-V sentences, and the sustained vowels and sentences combined.

3. Results

3.1. Question 1

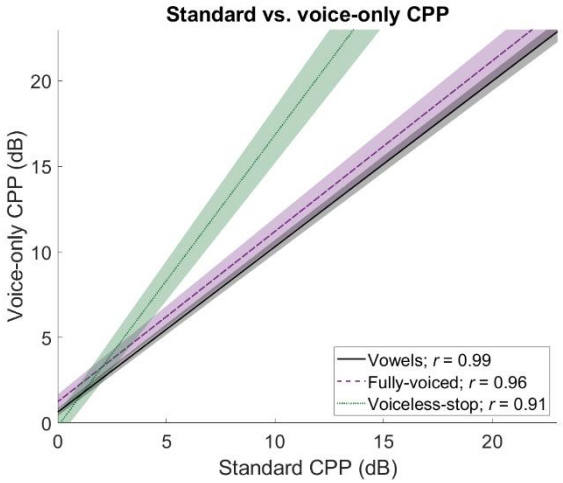

Figure 1: *Best-fit regression lines and 95% confidence intervals (shaded) for relationships between standard CPP (x-axis) and voice-only CPP (y-axis). Data is shown for vowels (black, solid line), the fully-voiced CAPE-V sentence (purple, dashed line) and the voiceless-stop CAPE-V sentence (green, dotted line).*

Figure 1 shows best-fit regression lines and 95% confidence intervals for relationships between standard CPP and voice-

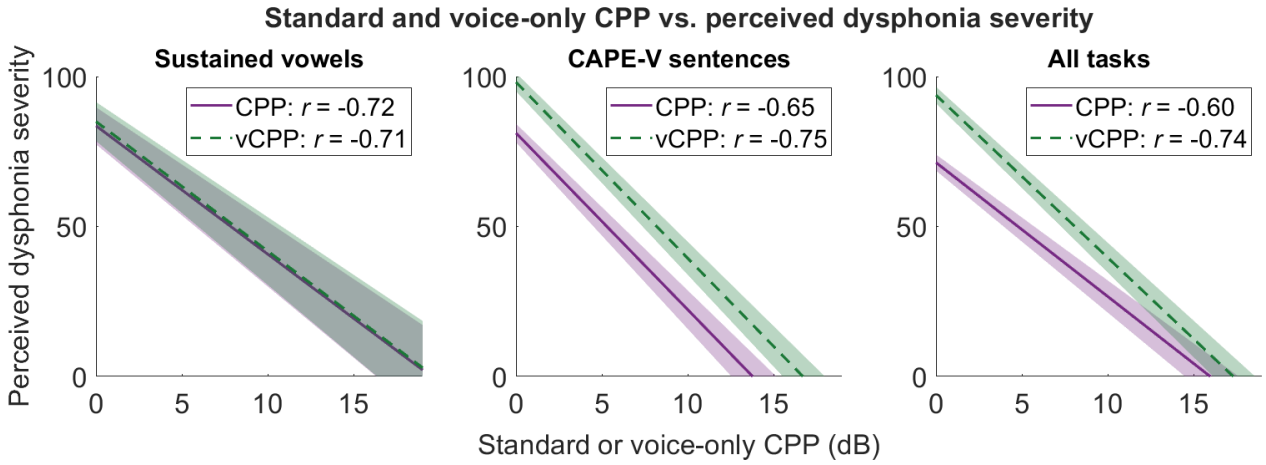

Figure 2: *Best-fit regression lines and 95% confidence intervals (shaded) for relationships between CPP (x-axis) and perceived dysphonia severity (y-axis). Data is shown for vowels (left), the fully-voiced CAPE-V sentence (center) and the voiceless-stop CAPE-V sentence (right). Solid purple lines show standard CPP and dashed green lines show voice-only CPP.*

only CPP. These relationships are shown for the sustained /a/ vowel, the fully-voiced CAPE-V sentence, and the voiceless-stop CAPE-V sentence separately. Pearson's correlation coefficient (r) was high for each relationship ($p < 0.05$). Additionally, the differences between all three pairs of correlation coefficients were significant using the R cocor package [20].

3.2. Questions 2 and 3

Table 1 summarizes the ROC AUCs and p-values used to assess Questions 2 and 3. For Question 2, classification is based on whether a speaker has a diagnosed voice disorder; for Question 3, classification is based on speakers' scores in the "Grade" section of the GRBAS. AUCs are presented for each of the six conditions described in the Methods section. Delong's test [21], as implemented in the R pROC package, was used to assess the significance of the difference in AUC between CPP and vCPP classification. The Bonferroni correction was applied to control for multiple comparisons.

Table 1: *ROC AUCs for classification accuracy based on standard and voice-only CPP in sustained vowels, CAPE-V sentences, and all tasks combined. P-values for the difference between the classification accuracy of standard and voice-only CPP for each group are also presented. * indicates comparisons with p < 0.0083 (i.e., 0.05/6 after Bonferroni correction).*

		CPP AUC	**vCPP AUC**	***p***
Voice disorder diagnosis (Question 2)	Vowels	0.710	0.699	0.012
	Sentences	0.747	0.767	0.049
	All tasks	0.718	0.752	**0.002***
Perceived dysphonia (Question 3)	Vowels	0.823	0.811	0.015
	Sentences	0.763	0.811	**0.001***
	All tasks	0.744	0.807	**0.001***

3.3. Question 4

Figure 2 shows best-fit regression lines and 95% confidence intervals for relationships between CPP (standard or voice-only) and perceived dysphonia severity. These relationships are shown based on data sets from the CAPE-V sentences (left) or from the full data set of CAPE-V sentences and sustained /a/ vowels (right). Pearson's correlation coefficient (r) was high for each relationship ($p < 0.05$). The differences between pairs of correlation coefficients were significant for tasks including continuous speech tasks but not for the sustained vowel condition. This finding indicates that perceived dysphonia severity correlated better with voice-only CPP than with standard CPP. Table 2 reports correlation coefficients and p-values for each task condition.

Table 2: *Pearson correlation coefficients for the relationships between perceived dysphonia severity and standard or voice-only CPP for vowels, CAPE-V sentences, and all tasks combined. P-values for the differences between standard and voice-only CPP are also reported. * indicates p < 0.05.*

	Standard CPP *r*	**Voice-only CPP *r***	***p***
Vowels	-0.72	-0.71	0.11
Sentences	-0.65	-0.75	**< 0.001***
All tasks	-0.60	-0.74	**< 0.001***

4. Discussion

This paper represents progress toward a novel method of computing CPP for voice analysis. Compared to traditional CPP computation, voice-only CPP is less affected by unvoiced content in the speech signal. In this study, we evaluated voice-only CPP in the context of speech with varying phoneme characteristics, including varying degrees of voiceless speech. However, we expect these results to extend to speech that varies in voicing content for other reasons, including alterations to speech rate and pausing. Our results address analytical and clinical validation of voice-only CPP following the V3 framework [16].

Analytical validation (Question 1) showed that voice-only CPP is less affected by unvoiced content in the speech signal. For sustained vowels, the very high correlation coefficient ($r \approx 0.99$) indicated that standard and voice-only CPP produce essentially equivalent results when there are no unvoiced regions to remove. Correlation coefficients were lower for tasks that were increasingly distinct from the sustained vowel task. Correlation coefficients between standard and voice-only CPP were slightly lower for the fully-voiced CAPE-V sentence ($r \approx 0.96$) compared to the sustained vowels, and lower again for the voiceless-stop-heavy CAPE-V sentence ($r \approx 0.91$). Overall, the relationship between standard and voice-only CPP is more vowel-like for continuous speech that is more fully voiced. This finding suggests that standard CPP values reflect the stimulus as well as the speaker's phonation, and that voice-only CPP computation is more robust to variation in tasks.

To address clinical validation, we asked how standard and voice-only CPP compare in identifying speakers with voice diagnoses (Question 2), identifying speakers with perceived dysphonia (Question 3), and in correlating with continuous judgments of perceived severity (Question 4).

Results for Question 2 indicated that standard and voice-only CPP had comparable accuracy in identifying speakers with diagnosed voice disorders within a single task type. In other words, with data based only on sustained /a/ vowels, or only on CAPE-V sentences, there was no difference in classification accuracy between the two CPP computation methods. However, when data from sustained vowels and CAPE-V sentences were combined, classification accuracy was higher when voice-only CPP was used compared to standard CPP. This finding comports with the intuition that voice-only CPP should provide a greater marginal benefit over standard CPP when data sets are more varied. Within a single task type, the range in CPP values is small enough that the added benefit of voice detection is relatively low. When task types are combined, variation in voice content creates enough variation in CPP values that voice detection is useful.

Question 3 concerned the ability of standard vs. voice-only CPP to identify speakers with perceptually dysphonic voices. This group overlaps, but is distinct from, the group of speakers with diagnosed voice disorders. Some speakers with mild or post-treatment voice disorders might have perceptually "normal" voices, and some speakers without a formal diagnosis might have undiagnosed voice disorders or temporary dysphonia for unrelated reasons (e.g., temporary illness or recent fatiguing voice use). For this classification, there was no difference in accuracy between standard and voice-only CPP for sustained vowels, but voice-only CPP was more accurate than standard CPP when continuous speech data was included. Both the CAPE-V-sentence-only conditions and the combination of CAPE-V sentences and sustained vowels showed increased classification accuracy with voice-only CPP.

Question 4 investigated the correlations between perceived severity (on a continuous 0-100 scale) and standard or voice-only CPP. Results indicated that this correlation was higher for voice-only CPP than for standard CPP for data that included continuous speech tasks. There was no difference between the correlation coefficients for the two CPP computation methods when sustained vowels were analyzed alone. Overall, these results suggest that the voice-only CPP algorithm presented here performs comparably to standard CPP for homogeneous data sets, especially those with fully-voiced speech stimuli, but that vCPP has several significant advantages over standard CPP for heterogenous data sets.

Note that the continuous speech tasks in this data set were all single sentences and generally did not contain many breaths or pauses. Any variation in voice percentage comes mainly from phoneme-level variation in the prevalence of unvoiced segments. We expect that the added benefit of voice-only CPP would be greater in data sets that include longer continuous speech utterances where breathing and pausing are more prevalent. Additionally, inter-speaker variation in pause rate and duration is expected to create even more variability in standard CPP values and further increase the benefit of voice detection.

5. References

[1] R. R. Patel *et al.*, "Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function," *American Journal of Speech Language Pathology*, vol. 27, no. 3, pp. 887–905, Aug. 2018, doi: 10.1044/2018_AJSLP-17-0009.

[2] O. Murton, R. Hillman, and D. Mehta, "Cepstral peak prominence values for clinical voice evaluation," *American Journal of Speech Language Pathology*, vol. 29, no. 3, pp. 1596–1607, Aug. 2020, doi: 10.1044/2020_AJSLP-20-00001.

[3] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, Nov. 2014, doi: 10.1016/j.bspc.2014.07.001.

[4] J. Delgado-Hernández, N. León-Gómez, and A. Jiménez-Álvarez, "Diagnostic accuracy of the Smoothed Cepstral Peak Prominence (CPPS) in the detection of dysphonia in the Spanish language," *loquens*, vol. 6, no. 1, p. 058, Feb. 2019, doi: 10.3989/loquens.2019.058.

[5] Y. W. Lee, G. H. Kim, I. H. Bae, H. J. Park, S. G. Wang, and S. B. Kwon, "The cut-off analysis using visual analogue scale and cepstral assessments on severity of voice disorder," *Logopedics Phoniatrics Vocology*, vol. 43, no. 4, pp. 175–180, Oct. 2018, doi: 10.1080/14015439.2018.1461925.

[6] M. Yu, S. H. Choi, C.-H. Choi, and B. Choi, "Predicting normal and pathological voice using a cepstral based acoustic index in sustained vowels versus connected speech," *Communication Sciences Disorders*, vol. 23, no. 4, pp. 1055–1064, Dec. 2018, doi: 10.12963/csd.18550.

[7] F. E. Aydinli, E. Özcebe, and Ö. İncebay, "Use of cepstral analysis for differentiating dysphonic from normal voices in children," *International Journal of Pediatric Otorhinolaryngology*, p. 7, 2019.

[8] S. N. Awan, N. Roy, D. Zhang, and S. M. Cohen, "Validation of the Cepstral Spectral Index of Dysphonia (CSID) as a screening tool for voice disorders: Development of clinical cutoff scores," *Journal of Voice*, vol. 30, no. 2, pp. 130–144, Mar. 2016, doi: 10.1016/j.jvoice.2015.04.009.

[9] C. R. Watts, S. N. Awan, and Y. Maryn, "A comparison of cepstral peak prominence measures from two acoustic analysis programs," *Journal of Voice*, vol. 31, no. 3, p. 387.e1-387.e10, May 2017, doi: 10.1016/j.jvoice.2016.09.012.

[10] G. B. Kempster, B. R. Gerratt, K. Verdolini Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," *American Journal of Speech Language Pathology*, vol. 18, no. 2, pp. 124–132, May 2009, doi: 10.1044/1058-0360(2008/08-0017).

[11] J. R. Green, D. Beukelman, and L. Ball, "Algorithmic estimation of pauses in extended speech samples of dysarthric and typical

speech," *Journal of Medical Speech-Language Pathology*, vol. 12, no. 4, p. 149, 2004.
[12] G. Fairbanks, "The Rainbow Passage," in *Voice and articluation drillbook*, 2nd ed., New York: Harper & Row, 1960, pp. 124–139.
[13] L. J. Platt, G. Andrews, and P. M. Howie, "Dysarthria of adult Cerebral Palsy," *Journal of Speech, Language, and Hearing Research*, vol. 23, no. 1, pp. 41–55, Mar. 1980, doi: 10.1044/jshr.2301.41.
[14] C. P. Shilpa, V. Swathi, V. Karjigi, K. S. Pavithra, and S. Sultana, "Landmark based modification to correct distortions in dysarthric speech," in *2016 Twenty Second National Conference on Communication (NCC)*, Mar. 2016, pp. 1–6. doi: 10.1109/NCC.2016.7561184.
[15] H. Kuwubara, "Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate," presented at the 5th European Conference on Speech Communication and Technology, Rhodes, Greece, Sep. 1997. [Online]. Available: https://www.isca-speech.org/archive_v0/archive_papers/eurospeech_1997/e97_1003.pdf
[16] J. C. Goldsack *et al.*, "Verification, analytical validation, and clinical validation (V3): The foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs)," *npj Digital Medicine*, vol. 3, no. 1, p. 55, Dec. 2020, doi: 10.1038/s41746-020-0260-4.
[17] P. Walden, "Perceptual Voice Qualities Database (PVQD)," vol. 3, Sep. 2020, doi: 10.17632/9dz247gnyb.3.
[18] M. Hirano, *Clinical Examination of Voice*. New York: Springer-Verlag, 1981.
[19] D. Hosmer, S. Lemeshow, and R. Sturdivant, *Applied logistic regression*, 3rd ed. Hoboken, New Jersey: John Wiley & Sons, Inc, 2013.
[20] B. Diedenhofen and J. Musch, "cocor: A comprehensive solution for the statistical comparison of correlations," *PLOS ONE*, vol. 10, no. 4, p. e0121945, Apr. 2015, doi: 10.1371/journal.pone.0121945.
[21] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988, doi: 10.2307/2531595.