Stability of long-term average spectrum measures and cepstral peak prominence in connected speech

Jörg Peters

University of Oldenburg

joerg.peters@uol.de

Abstract

In order to compensate for the segmental variability of connected speech, spectral measures derived from the long-term average spectrum (LTAS) are usually calculated from speech samples lasting more than 20 seconds. The present paper examines whether a shorter measurement interval is sufficient to compensate for segmental variation, and whether the integration time of spectral measures differs from that of cepstral measures, such as the smoothed cepstral peak prominence (CPPS). In addition, the influence of speech type and language on the integration time of these measures is examined. Speech samples were recorded from 61 bilingual speakers of High and Low German. Each speaker read two passages, a coherent narrative and a non-coherent list of sentences, in both languages. Spectral slope measures, the speaker's formant, and CPPS were examined. The integration time was found to vary substantially among these measures. CPPS had the shortest integration time and the slope measures the longest. The slope measures were also most strongly affected by speech type and language.

Introduction

The long-term average spectrum (LTAS) is one of the most common research tools in the acoustic analysis of voice quality. While the study of the singing voice and of voice disorders is usually based on sustained vowels, there is a growing number of phonetic studies calculating LTAS measures from connected speech. To account for the segmental variability of connected speech, LTAS measures are usually taken from longer speech samples ranging from 20 to 90 seconds (e.g. Blomberg, & Elenius 1970, Fritzell, Hallen, & Sundberg 1974, Doherty 1975, Hollien, & Majewski 1977, Hammarberg et al. 1980, Boves 1984, Weiss 1993, Byrne et al. 1994, Engelbert 2014, Bahmanbiglu, Mojiri, & Abnavi 2017). According to Majewski, Rothman, & Hollien (1977), one minute of speech is required to get stable measurements. Mendoza et al. (1996) observe stable LTAS signals for most of their speakers at 25 seconds. For average spectra calculated from intervals differing by about 5 seconds, Coadou and Rougab (2007) report cross-correlation coefficients which are above 0.999 after 29 seconds.

The present study examines whether a shorter measurement interval is sufficient, and whether measures derived from the LTAS differ by the time they need to stabilize. Shorter integration times would be of interest not only for forensic studies but also for the study of short- and mid-term changes of voice quality during single speech tasks.

Two spectral slope measures and the speaker's formant are examined in read speech. Slope measures correlate with phonatory mode (Stevens & Hanson 1995). The speaker's formant, which is characterized by an energy concentration at 3.5 kHz in male speakers, results from increased resonance in the vocal tract (Acker 1987, Leino 1993). We also consider the smoothed cepstral peak prominence (CPPS), which indicates the amount of acoustic energy transmitted by the harmonic components of the speech signal (Hillenbrand & Houde 1996). As CPPS is expected to be less affected by segmental variation of connected speech it is often calculated from a single sentence (e.g. Hillenbrand, & Houde 1996, Sauder, Bretl, & Eadie 2017).

To examine possible effects of speech type (e.g. Keller 2003), we compare readings of a coherent and a non-coherent text. To examine possible language effects (e.g. Bahmanbiglu et al. 2017), we compare speech samples recorded from bilingual speakers in two language modes.

Method

The analysis is based on recordings of 61 subjects (34 males, 27 females), aged 40–82. All subjects were born in the *Bersenbrücker Land* in northwest Germany and they were bilingual with Standard High German and the local variant of Low German. Each subject read aloud two passages, the fable “The North Wind and the Sun” (= Narrative) and the 40 sentences, which were conceived by Georg Wenker for a dialectological survey in the late 19th century known as the *Sprachatlas des Deutschen Reichs* (= Wenker sentences). Whereas the narrative forms a coherent text, the Wenker sentences form a non-coherent text consisting of a list of unrelated sentences.

The speech samples were recorded with a DAT recorder (TASCAM DR-100 MK3) with a sampling rate of 48 kHz at 24 bit resolution and were later resampled to 16 kHz. With the help of a *Praat* script (Boersma & Weenink 2017) provided by Wilbert Heeringa, acoustic pauses and unvoiced segments were automatically removed (for an alternative method using vowel nucleus detection see Keller 2003). Three spectral measures and one cepstral measure were calculated in *Praat*: (i) *Slope I,* the energy level difference between 1–5 kHz and 0–1 kHz (Frokjaer-Jensen, & Prytz 1976, Löfqvist, & Mandersson 1987), (ii) *Slope II,* the energy level difference between 5–8 kHz and 1–5 kHz (Löfqvist, & Mandersson 1987, Guzman et al. 2013), (iii) *SF,* the speaker’s formant, which was calculated from the difference between the energy level at 3–3.8 kHz and the adjacent energy levels at 2.2–3 kHz and 3.8–4.6 kHz following Boersma and Kovacic (2006), and (iv) *CPPS*, which was calculated with the settings used by Barsties and Maryn (2016) for the calculation of the Acoustic Voice Quality Index.

To determine the integration times of the spectral and cepstral measures, two distance measures were calculated. First, we measured the distance (in dB) of the cumulative average from the average obtained from a reference interval, which was set to 100 seconds (= *Distance I).* Second, we calculated the change of the cumulative average (in dB) that results from extending the measurement interval by one second *(= Distance II).* Both measurements were taken every 0.2 seconds.

Results

Figure 1 shows the distances of the cumulative averages for Slope I, Slope II, SF, and CPPS from the averages obtained for the 100 second interval (*Distance I).* These measurements were taken from the Wenker sentences only as the recordings of the narrative were shorter than 100 seconds.

CPPS has the shortest integration time. After about two seconds, the cumulative average approaches the reference value of the 100 second interval to less than 0.2 dB. SF reaches this distance after about 20 seconds and Slope I after about 40 seconds. The longest integration time was found for Slope II. It approaches the reference value to less than 0.2 dB after about 60 seconds. On the other hand, the volatility of all spectral measures declined strongly after 20–25 seconds, followed by a trending phase towards the final level during the following 40–50 seconds. The decline of the volatility can even more clearly be observed in Figure 2, which shows the change of the cumulative averages by extending the measurement interval by one second (*Distance II).*

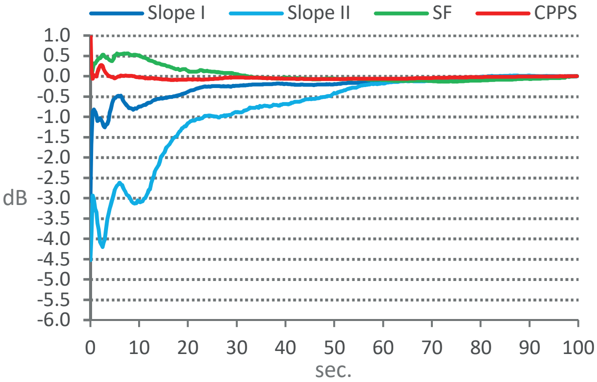

Figure 1. Distance I for Wenker sentences (N = 61).

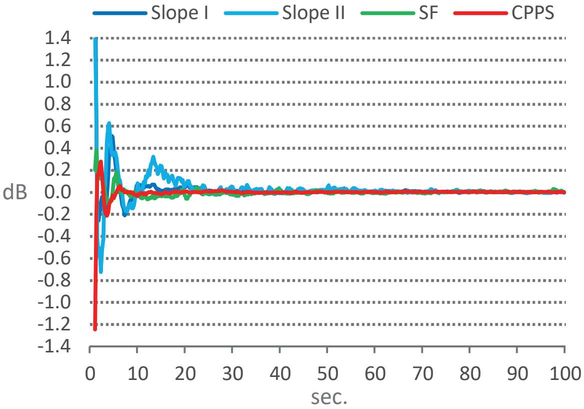

Figure 2. Distance II for Wenker sentences (N = 61).

Figures 3 and 4 illustrate the effects of LANGUAGE and SPEECH TYPE on *Distance II*. The results suggest effects of both factors on the slope measures. Whereas the effect of LANGUAGE differs for Slope I and Slope II, there is a uniform effect of SPEECH TYPE. For both Slope I and Slope II the non-coherent text was found to have a longer integration time than the coherent text. No substantial effects of LANGUAGE and SPEECH TYPE were found for SF and CPPS. In all conditions, the cumulative averages of SF and CPPS changed by less than 0.2 dB after 2–3 seconds.

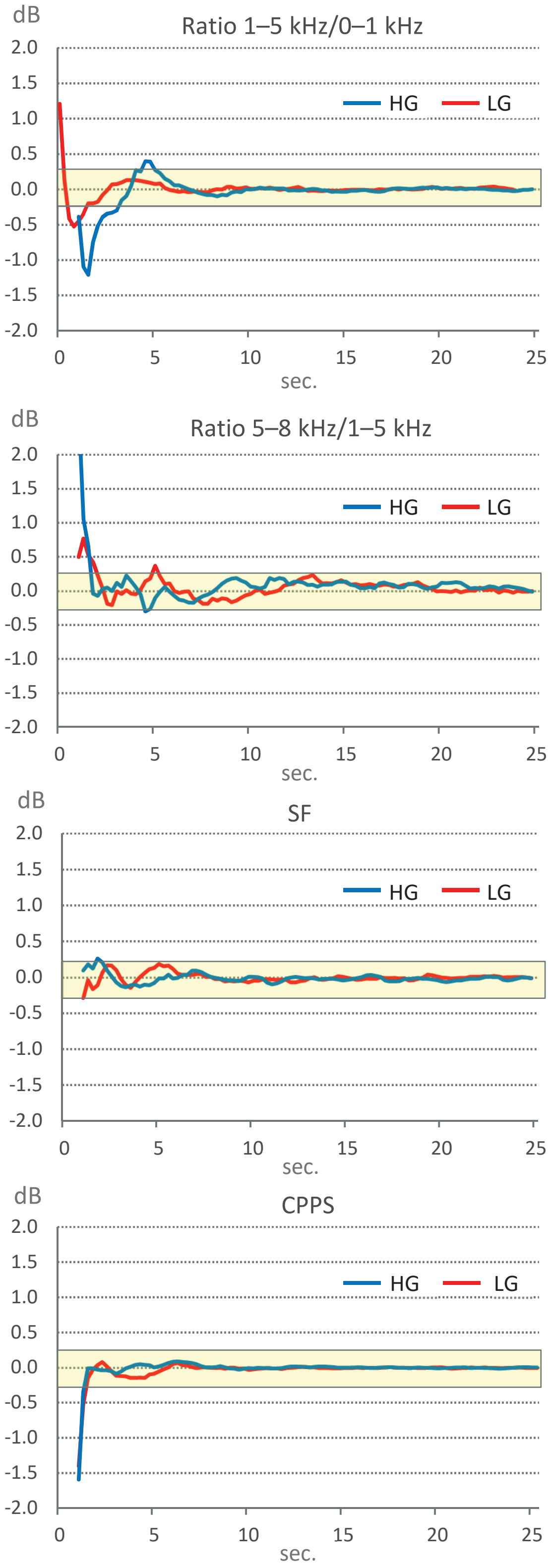

Figure 3. Distance II for Low German and High German recordings (N = 244). The channel indicates a change of ± 0.2 dB.

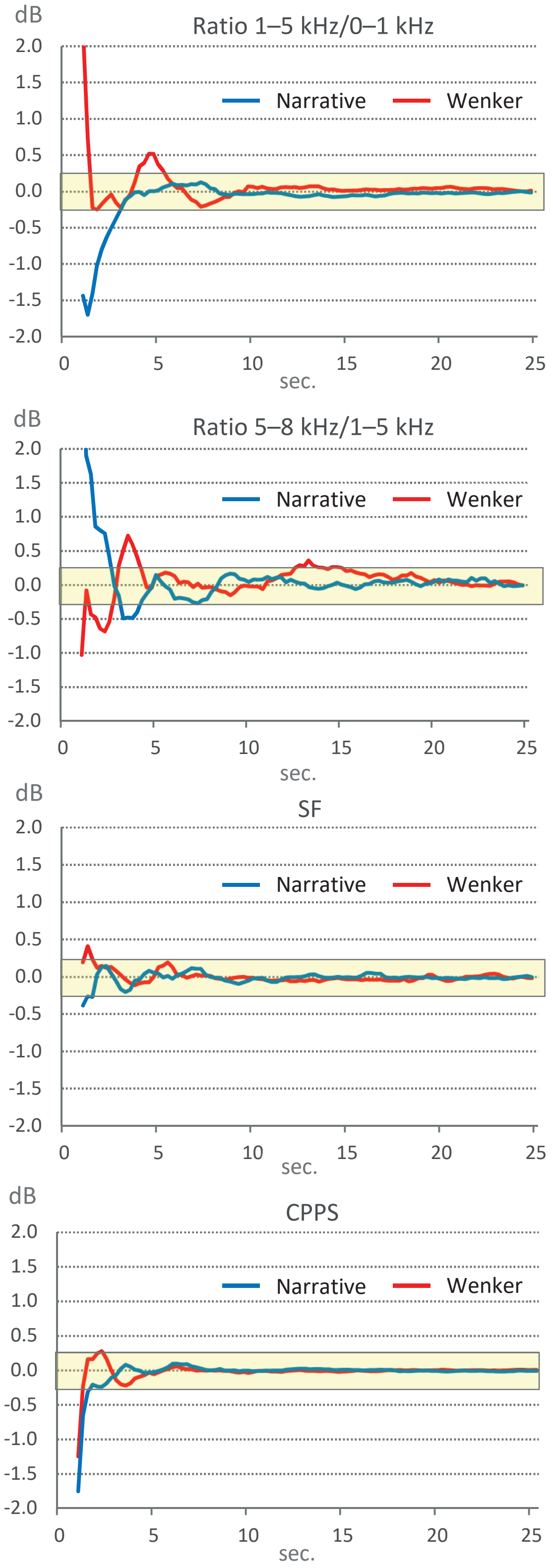

Figure 4. Distance II for recordings of the narrative and the Wenker sentences (N = 244). The channel indicates a change of ± 0.2 dB.

Discussion

Integration time was found to vary among different acoustic measures of voice quality and to be affected by language and speech type. The spectral slope measure Slope II, which includes acoustic energy above 5 kHz, required most time to stabilize, and CPPS and SF the least. LANGUAGE (Low German vs. High German) and SPEECH TYPE (coherent vs. non-coherent text) mainly affected the spectral slope measures. Our results suggest that the interval sizes of

20–40 seconds, which are regarded as sufficient by most researchers, are reasonable for obtaining stable slope measures. On the other hand, SF, like CPPS, needs less than 5 seconds of connected speech to reduce the volatility to a very low level. These results suggest that at least some acoustic measures may be suitable for examining short- to mid-term changes of voice quality during speech tasks. To study the dynamics of SF and CPPS, a moving average with an interval size of 2–5 seconds may be sufficient. Note, however, that the integration times calculated in the present study are based on group means, where positive and negative fluctuations of individual measurements balance each other out. For individual performances and for small speaker groups longer integration times are to be expected.

References

Acker, B. F. (1987). Vocal tract adjustments for the projected voice. *Journal of Voice,* 1, 77–82.

Bahmanbiglu, S. A., Mojiri, F., & Abnavi, F. (2017). The impact of language on voice. An LTAS study. *Journal of Voice* 31, 249.e9–249.e12.

Barsties, B., & Maryn, Y. (2016). External validation of the Acoustic Voice Quality Index version 03.01. With extended representativity. *The Annals of Otology, Rhinology, and Laryngology,* 125, 571–583.

Blomberg, M., & Elenius, K. (1970). Statistical analysis of speech signals. *STL–QPSR,* 11, 1–8.

Boersma, P., & Kovacic, G. (2006). Spectral characteristics of three styles of Croatian folk singing. *The Journal of the Acoustical Society of America,* 119, 1805–1816.

Boersma, P., & Weening, D. (2017). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.31, retrieved 22 August 2017 from http://www.praat.org/

Boves, L. W. J. (1984). *The phonetic basis of perceptual ratings of running speech*. Dordrecht etc.: Foris Publications.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., et al. (1994). An international comparison of long-term average speech spectra. *The Journal of the Acoustical Society of America,* 96, 2108–2120.

Coadou, M., & Rougab, A. (2007). Voice quality and variation in English. *Proceedings of the 16th International Congress of Phonetic Sciences, XVI, 6 – 10 August 2007, Saarbrücken, Germany*. Saarbrücken: Universität des Saarlandes, 2077–2080.

Doherty, E. T. (1975). *An evaluation of selected acoustic parameters for use in speaker identification*. Diss. University of Florida.

Engelbert, A. P. (2014). Cross-linguistic effects on voice quality: A study on Brazilians' production of Portuguese and English. *Proceedings of the International Symposium on the Acquisition of Second Language Speech. Concordia Working Papers in Applied Linguistics*, 5, 157–170.

Fritzell, B., Hallen, O., & Sundberg, J. (1974). Evaluation of teflon injection therapy for paralytic dysphonia. *STL–QPSR,* 15, 18–24.

Frokjaer-Jensen, B., & Prytz, S. (1976). Registration of voice quality. *Bruel and Kjaer Technical Review,* 3, 3–17.

Guzman, M., Correa, S., Muñoz, D., & Mayerhoff, R. (2013). Influence on spectral energy distribution of emotional expression. *Journal of Voice,* 27, 129e1–129e10.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Oto-Laryngologica,* 90, 441–451.

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech and Hearing Research* 39, 311–321.

Hollien, H., & Majewski, W. (1977). Speaker identification by long-term spectra under normal and distorted speech conditions. *The Journal of the Acoustical Society of America,* 62, 975–980.

Leino, T. (1993). Long-term average spectrum study on speaking voice quality in male actors. *Proceedings of the Stockholm Music Acoustics Conference*, vol. 28. The Royal Swedish Academy of Music Stockholm.

Löfqvist, A., & Mandersson, B. (1987). Long-time average spectrum of speech and voice analysis. *Folia Phoniatrica,* 39, 221–229.

Keller, E. (2003). Voice characteristics of MARSEC speakers. *VOQUAL'03*, Geneva, August 27–29, 2003.

Majewski, W., Rothman, H., & Hollien, H. (1977). Acoustic comparisons of American English and Polish. *Journal of Phonetics,* 5, 247–251.

Mendoza, E., Muñoz, J., Valencia Naranjo, N. (1996). The long-term average spectrum as a measure of voice stability. *Folia Phoniatrica et Logopaedica,* 48, 57–64.

Sauder, C., Bretl, M., & Eadie, T. (2017). Predicting voice disorder status from smoothed measures of cepstral peak prominence using *Praat* and *Analysis of Dysphonia in Speech and Voice* (ADSV). *Journal of Voice*, 31, 557–566.

Stevens, K. N., & Hanson, H. M. (1995). Classification of glottal vibration from acoustic measurements. In: O. Fujimura, & M. Hirano (eds.), *Vocal fold physiology: Voice quality control*. San Diego, Calif.: Singular Publ., Inc., 147–170.

Weiss, W. (1993). Comparison of spectral and voice quality parameters of voice trained Anglophones and Francophones in continuous speech. *Canadian Acoustics,* 21, 3–8.