Shouting affects temporal properties of the speech amplitude envelope

Dimos, Kostis ; He, Lei ; Dellwo, Volker

DOI: https://doi.org/10.1121/10.0023995

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Shouting affects temporal properties of the speech amplitude envelope

Kostis Dimos,[a] **Lei He, and Volker Dellwo**
Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
kostis.dimos@uzh.ch, lei.he@uzh.ch, volker.dellwo@uzh.ch

Abstract: Distinguishing shouted from non-shouted speech is crucial in communication. We examined how shouting affects temporal properties of the amplitude envelope (ENV) in a total of 720 sentences read by 18 Swiss German speakers in normal and shouted modes; shouting was characterised by maintaining sound pressure levels of ≥80 dB sound pressure level (dB-SPL) (C-weighted) at a 1-meter distance from the mouth. Generalized additive models revealed significant temporal alterations of ENV in shouted speech, marked by steeper ascent, delayed peak, and extended high levels. These findings offer potential cues for identifying shouting, particularly useful when fine-structure and dynamic range cues are absent, for example, in cochlear implant users.

[Editor: Douglas D. O'Shaughnessy] https://doi.org/10.1121/10.0023995

Received: 20 June 2023 **Accepted:** 27 November 2023 **Published Online:** 3 January 2024

1. Introduction

Shouting, driven by heightened vocal effort, manifests in diverse contexts and for various purposes. It serves to amplify speech signals for extended distances (Traunmüller and Eriksson, 2000) or can emerge in speech associated with intense emotions, such as anger, fear, and aggression. Shouting is often prompted by alarming or threatening situations (Gangamohan *et al.*, 2019; Mittal and Yegnanarayana, 2013), or it can serve as a means of conveying urgency (Jang, 2007; Kobayashi *et al.*, 2022).

The ability to detect shouted speech styles is essential for effective social communication. Understanding the acoustic cues to shouted speech is paramount in this regard. Among these cues, average intensity emerges as a prominent indicator. Crucially, however, shouted speech typically remains identifiable when replayed at lower average intensity. Hence, there must be other cues by which shouted speech can be identified. For example, louder speech levels are associated with an increased vocal fold vibration rate, resulting in the perception of higher pitch (Mittal and Yegnanarayana, 2013). Increased lip opening and lowered jaw and tongue in shouted speech (Xue *et al.*, 2021) in connection with higher subglottal pressure and muscle tension not only lead to increased sound pressure levels (SPLs), but also lead to increased fundamental frequency (f_o) and f_1 in shouting. There are also differences in relative energy levels, with high energy being distributed predominantly in voiced intervals of shouted speech and less in voiceless intervals (Baghel *et al.*, 2021; Mittal and Vuppala, 2016a). Additionally, elevated f_o and dynamic ranges (Baghel *et al.*, 2021; Bonnot and Chevrie-Muller, 1991; Jang, 2007; Kobayashi *et al.*, 2022; Mittal and Vuppala, 2016b; Raitio *et al.*, 2013; Xue *et al.*, 2021; Zhang and Hansen, 2007) along with a shift in the energy distribution across frequency bands, with increased energy in low frequencies (Pohjalainen *et al.*, 2013; Ternström *et al.*, 2006; Zhang and Hansen, 2007), are demonstrated cues to shouting. Especially a quick f_o rise time in shouted speech has been shown to have a strong impact on perceived urgency (Jang, 2007; Kobayashi *et al.*, 2022), even though these effects can be language specific (Kobayashi *et al.*, 2022). Notably, there are differences between gender in the production and detection of shouted speech related to f_o (Baghel *et al.*, 2021): for example, an overlap between male shouted and female normal average f_o, caused by changes in the excitation source in male shouted speech. Higher SPLs in male compared to female speakers were reported in Lombard speech Ternström *et al.* (2006) and also higher vowel duration for the female speakers (Alghamdi *et al.*, 2018). An increase in f_1 has been found in high vocal effort speech (Liénard and Benedetto, 1999; Pohjalainen *et al.*, 2013; Traunmüller and Eriksson, 2000; Xue *et al.*, 2021), most likely standing in relation to shortening of the pharyngeal cavity during shouting as an effect of muscular tension in the back of the vocal tract. Schulman (1989) and Pohjalainen *et al.* (2013) pointed out a relationship between f_o and f_1 in shouted speech based on scaling regarding vowel quality and maintaining Bark distance between the two acoustic correlates. These effects could not be replicated for f_o and the higher vowel formants (Liénard and Benedetto, 1999).

[a] Author to whom correspondence should be addressed.

In summary, all previously discussed cues to shouted speech are related to either average intensity differences between normal and shouted speech, differences in the dynamic range, or differences in the temporal fine-structure of speech (f_o or formant frequencies). In this study, we wanted to understand whether the temporal organisation of speech varies between normally read and shouted sentence utterances. In the temporal domain, vowels tend to increase in duration, while consonants typically become shorter (Dromey *et al.*, 1995; Raitio *et al.*, 2013; Rostolland, 1982; Schulman, 1989), resulting in a marginal increase of sentence and CV intervals duration (Schulman, 1989; Zhang and Hansen, 2007). However, measurements of interval duration alone do not offer comprehensive insights into underlying effects within these intervals. Here, we therefore investigated the effects of shouted speech on the temporal development of the amplitude envelope (ENV) over the course of the utterance. Specifically, we analysed how ENV is modulated within a voiced sequence of speech. We approached this by modelling ENV in voiced intervals of normal and shouted utterances using generalized additive models (GAMs). Our specific aim was to understand a possible relationship between the increased duration of voiced speech intervals and the trajectory of ENV. Such cues may be particularly relevant in situations in which fine-structure cues like f_o or formants and dynamic range cues are limited, for example, in cochlear implant (CI) users (see Sec. 3).

Shouted speech consists of a large variety of forms of high vocal effort speech, and thus, obtaining comparable varieties of shouted speech from speakers is not necessarily trivial. Cushing *et al.* (2011) categorized methods into two primary groups: those focused on perceived speech level, measured in terms of loudness, and those centered on the speakers and their vocal effort during production. Vocal effort refers to the exertion perceived or reported by speakers (Baldner *et al.*, 2015; McKenna and Stepp, 2018), making it a subjective physiological measure (Traunmüller and Eriksson, 2000) distinct from SPL based measurements. Cushing *et al.* (2011) and Zhang and Hansen (2007) consider shouted speech as the highest point in a range of five levels, starting from "hushed" (Cushing *et al.*, 2011) or whispered (Zhang and Hansen, 2007) speech. In shouted speech, Cushing *et al.* (2011) have measured averages of 95 A-weighted decibels [dB(A)] and 88 dB(A) at a distance of 50 cm for males and females, respectively. Zhang and Hansen (2007) reported results from male speakers, measured at a distance of 75 cm, with the intensity values ranging between 75 and 90 dB-SPL approximately. Here, we asked listeners to shout sentence utterances at a minimum of 80 dB-SPL (C-weighted) measured from a 1-meter distance in a closed recording booth with high damping. Even though this target was reached in varying ways across different subjects, all speakers produced canonical variants of shouting via this setup.

2. Method

2.1 Participants

Eighteen native Swiss-German speakers (9 males/9 females) were recruited from the student population at the University of Zurich to participate in the experiment. The participants gave their informed consent to participate in the study and were paid for their participation. Before every recording session, each speaker completed a screening questionnaire, "Voice Handicap Index-9 international" (VHI-9i) (Nawka *et al.*, 2009), to confirm that his/her voice is healthy and would not be affected after the recordings (Brockmann-Bauser and Bohlender, 2015).

2.2 Recording procedure

A set of 20 semantically neutral sentences, each consisting of an equal number of words and sharing identical grammatical structures, was created based on the Oldenberg Sentence Test (OLSA) model (Wagener *et al.*, 1999). Each sentence was presented to the speakers in written on a monitor inside a recording booth. Each recording session began with the normal speech condition, during which speakers read the sentences displayed on the screen in a way they considered reading at normal vocal effort. Subsequently, speakers read the same sentences out loud with a break after the first 10 shouted utterances. Recordings took place in a noise-controlled room at the University of Zurich. K.D. and L.H. carried out the recordings and instructed participants to produce shouted speech. Speakers were recorded with a headset in which the microphone was placed about 10 cm from the speaker's mouth. Sound pressure levels were measured on a C-weighted, fast response dB-SPL meter at a 1 m distance from the speaker's mouth. A threshold of 80 dB-SPL had to be reached (Baken and Orlikoff, 2000). Participants were provided visual feedback on a screen, which turned green when 80 dB-SPL for the shouted utterance was reached. Failing to make the screen turn green resulted in repeating the sentence utterance. To calibrate recording input-levels, the speaker produced a prolonged [*a* :] at what the speaker regarded a normal vocalisation effort. Input level was set to roughly −6 decibel volume units (dBVU) for this vocalisation. Under the shouted condition, another [*a* :] was produced by the speaker that had to reach 80 dB-SPL in a 1 meter distance. Again, for this production, the input level was set to roughly −6 dBVU. This procedure resulted in roughly equal mean intensity values for normal and shouted utterances. For each speaker, the shouted speech recording was conducted in two parts of 10 utterances each, with a break in between the two sessions. We consider the use of the term "normal," in contrast to "shouted," to be consistent with the previous studies on the topic (Baghel *et al.*, 2021; Bonnot and Chevrie-Muller, 1991; Jang, 2007; Mittal and Vuppala, 2016a; Pohjalainen *et al.*, 2013; Raitio *et al.*, 2013; Schulman, 1989; Xue *et al.*, 2021; Zhang and Hansen, 2007).

2.3 Acoustic analysis

The utterances were analysed into voiced and unvoiced intervals using PRAAT's auto-correlation function (pitch range, 75 to 600 Hz; maximum period, 20 ms; mean period, 10 ms) (Boersma and Weenick, 2022). ENV was obtained by calculating an intensity contour with the software PRAAT. Intensity contours were calculated by taking the root-mean-square (rms) value of a 32 ms window with a forward at 8 ms intervals, resulting in 125 intensity values/s. The 32 ms window duration produces roughly equal ENV contours to common comparable methods based on low-pass filtering of a Hilbert signal or a full-wave rectified signal at 16 Hz. Intensity contours were in dB-SPL; however, the reference value equating 0 dB-SPL was that of a sinusoid with an amplitude of 0.000 028 284 27 ($\sqrt{2 \cdot \text{rms}^2}$); with rms = 0.000 02 and not a 20 μPa sinusoid in air. Thus, the absolute dB-SPL values were meaningless. The input level was adjusted for the recording not to clip (cf. Sec. 2.2). Lagier *et al.* (2017) show that non-periodic voice may appear in long shouts that reach maximal dB-SPL, a manual review of the voiced interval annotations was performed after the data analysis to ensure no such effects would exist in a way that would significantly impact our statistical analysis. To carry out a GAM analysis, the intensity contour was normalized in time for each voiced interval by resampling the interval into 20 equidistant sampling points (henceforth referred to as "timesteps"). The intensity values were *z*-normalized for each speech condition (normal and shouted) to eliminate absolute intensity differences between normal and shouted speech.

3. Results and discussion

Figure 1 shows an increase in the average duration in voiced intervals. Additionally, the absolute number of voiced intervals is higher in the normal compared to the shouted condition [Fig. 1(b)]. All effects were highly significant [linear mixed effect models (Winter, 2013)] with random intercepts for speaker, sentence and gender cf. results of likelihood ratio tests between the full and reduced models for each fixed effect in Table 1). These findings are consistent with previous findings of increased voiced interval duration in shouted speech (Dromey *et al.*, 1995; Raitio *et al.*, 2013; Rostolland, 1982; Schulman, 1989). The median duration of voiced intervals increased from approximately 150 ms to 220 ms in shouted speech, with interquartile ranges (IQRs) between 128 and 190 ms in normal speech and 185 and 225 ms in shouted speech. Finally, the median for number of voiced intervals per utterance was reduced from 7 to 6 in shouted speech, with an IQR between 6 and 9 intervals in normal speech and an IQR between 5 and 8 in shouted speech. Mean intensity of each utterance was on average about 5 dB-SPL higher in shouted compared to normal voices. This was because the adjustment of input level did not compensate fully for the intensity differences between shouting and normal speech. Given that all ENV contours were normalized (*z*-score), the absolute differences in intensity between utterances did not play a role in further analysis.

The GAM (Wood, 2011, 2017,) was fitted using restricted maximum likelihood (REML) estimation to analyze the impact of shouted speech on ENV (Table 2). The timestep parameter, comprising the sampled values within each

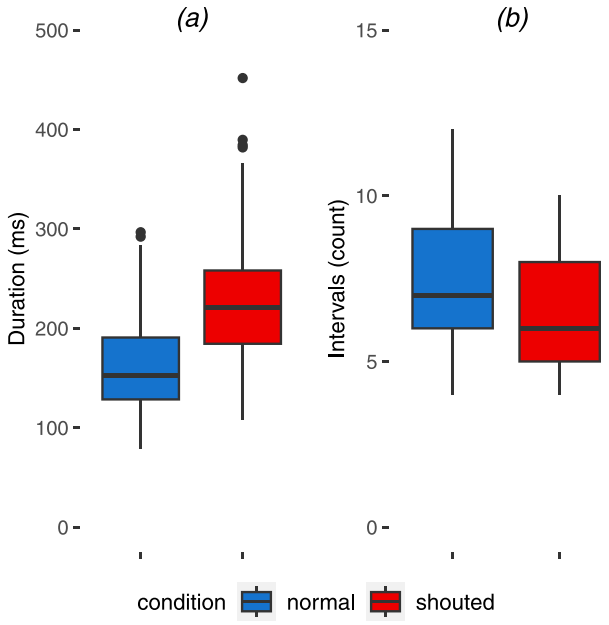

Fig. 1. Boxplots on voiced intervals showing the distributions of utterance averaged (a) interval duration and (b) the number of voiced intervals per utterance across the two conditions.

Table 1. Results of likelihood ratio tests for model comparisons between the full and reduced models.[a]

Parameter	df	AIC	BIC	Log likelihood ratio	Deviance	χ^2 (df)	P value
Duration	8	−2867.7	−2831.1	1441.8	−2883.7	36.26 (1)	<0.000
Intervals	8	2030.3	2066.9	−1007.1	2014.3	26.027 (1)	<0.000

[a]AIC, Akaike information criterion; BIC, Bayesian information criterion.

interval, was included through the use of a cubic spline. Speaker and sentence variables were introduced as random effects.

As Table 3 shows, the condition parameter had a significant effect on the normalized ENV values ($P = 0.008\,28$). Similarly, the timestep parameter and the two random effects, namely speaker and sentence, were found to be significant ($P < 0.000$), as well as gender ($P = 0.002\,23$), indicating that these parameters are also significant contributors to the variability in the ENV contour. Overall, the model explains approximately 38% of the variability in the ENV values adjusted ($\chi^2 = 0.378$).

Figure 2(a) shows the two normalized ENVs of the voiced intervals, averaged over speakers. The values are aggregated over all 20 sentences across all speakers. In the beginning of the interval, shouted speech is characterised by a much steeper ascent of the contour than normal speech. This steeper climb of ENV might be explained by the increased level of intensity speech eventually reaches during the interval. This appears to be the case for most of the speakers, although for speakers 2, 3, and 12, the difference is smaller. The ascent extends for a longer time, crucially leading to a delayed peak in shouted speech—two sample points later in these averaged contours. We should perhaps consider how the longer ascent and the delayed peak, taken together, constitute manifestations of the increased vocal effort applied by the participants for the production of shouted speech. We assume that the increased vocal effort and the amplified articulatory movements in the shouted condition exert an impact on the temporal structure of ENV. Therefore, we assert that the time distance between the onset of a voiced interval and the point that the ENV peak is reached is longer when the reorganization of articulatory movements required in shouting takes place. We assume that this may be a salient feature of shouted speech that may be detected in the absence of fine-structure cues.

In a normal speech contour, there is a steady decline after the peak, whereas in shouted speech, there is a high-level plateau, or in some cases a secondary peak, as indicated for speakers 4, 5, 6, 11, and 14. We will consider two possible interpretations for this observation. First we have to take into account that the voiced intervals in shouted speech have a longer duration and that the acoustic information mostly relevant for the perception of shouted speech lies in these high-energy, voiced intervals (Baghel *et al.*, 2021; Mittal and Vuppala, 2016a). Therefore, speakers may be attempting to maintain a high level of intensity during these intervals. A different point of view would be to consider the fact that there are fewer intervals per utterance in shouted speech. As we have seen earlier, Fig. 1 indicates a lower number of voiced intervals in shouted speech. It is conceivable that this decrease is the result of some of those intervals being merged, resulting in prolonged high intensity and secondary peaks. Merged intervals may occur across word boundaries or as an indirect effect of voiceless consonants' deletion or underproduction. Alternatively, unvoiced intervals may become voiced in shouted speech, resulting in fewer, but longer, continuous, voiced intervals per utterance.

Between-speaker variability can be observed in the individual contour figures in Fig. 2(b), which shows ENV per speaker aggregated over all the 20 sentences. The speaker and gender parameters had a significant effect on the variability of ENV across the conditions. Speaker individual characteristics in articulation, speech rate, and loudness may affect the levels in which they have to adapt or the amount of vocal effort they need to apply in order to reach the required sound pressure threshold during the experiment. Further analysis and experimentation may reveal speaker-specific characteristics that can better explain the differences in the effect of shouting on ENV.

Understanding the effect of shouting on detailed temporal adjustments and the re-organization of ENV and f_o contours can shed light on how listeners may potentially identify shouted speech: for example, when it is replayed at low average intensity. The temporal organisation of the shouted ENV may particularly be a potential cue to shouting in the

Table 2. Generalized additive model setup and parameters.

Parameter	Smooth terms	Description
ENV	Dependent variable	z-normalized
Condition	Fixed factor	Normal, shouted
Timestep	Cubic spline	Sample location in the time-domain (1–20)
Speaker	Random factor	Random effect parameter for the 18 speakers
Gender	Random factor	Random effects of gender: male, female
Sentence	Random factor	Random effect parameter for sentences

Table 3. GAM results that show a significant effect of speech condition (normal, shouted) on the ENV of voiced intervals. The other parameters were also significant.

Parameter		Estimate	Standard error	Pr(>|t|)
(Intercept)		−0.000 971 6		
Condition:shouted ratio		0.013 128 3	0.004 972 0	0.008 28
Parameter	**edf[a]**	**Reference df**	***F***	***P* value**
Timestep	8.894	8.996	2026.32	<0.000 1
Speaker	16.993	17	3508.35	<0.000 1
Gender	0.000 275 9	1	3.743	0.002 23
Sentence	18.157	19	41.28	<0.000 1

[a]edf, effective degrees of freedom.

absence of fine-structure and dynamic range cues. Such situations can be found in noise-vocoded speech and in CI users who lack the perception of f_o and have strongly reduced dynamic ranges (Adel *et al.*, 2019; Meister *et al.*, 2011), in particular in high-frequency tones (Tak and Yathiraj, 2019), as well as limitations in perceiving higher levels of pitch (Kong and Carlyon, 2010) and a correlation between intensity variations and perceived pitch levels (Arnoldner *et al.*, 2006). In return, CI users can accurately detect and identify syllable duration variations (Meister *et al.*, 2011) and they rely strongly on information in ENV in the interpretation of speech cues (Fischer *et al.*, 2021; Wilson *et al.*, 1991). It will be interesting to

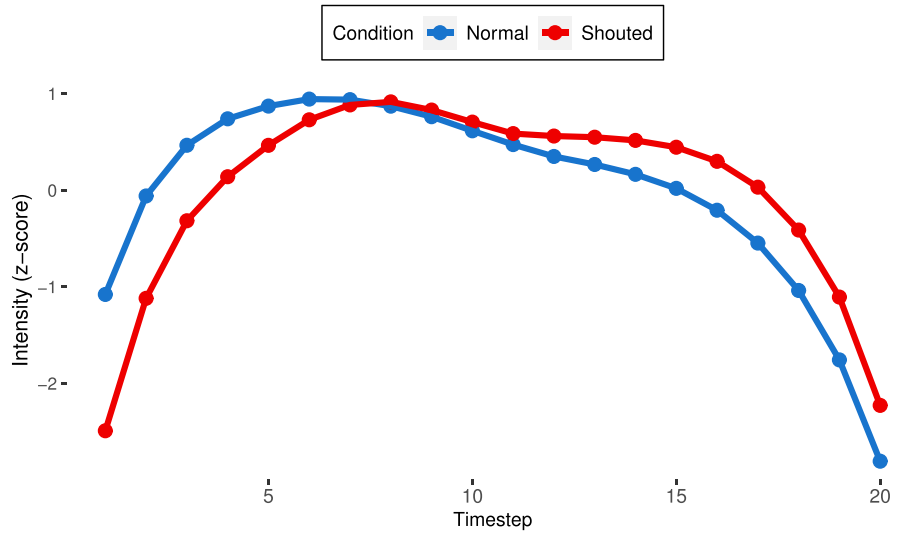

Fig. 2. ENV of voiced intervals in normal and shouted speech. (a) Mean aggregated by sample. (b) Mean aggregated by individual speakers.

test to what degree cochlea implant users can distinguish normal from shouted speech in general and to what degree ENV characteristics play a role in this process. Similarly, it might be interesting to understand whether these cues play a role in the processing of shouted speech in normal listeners. Furthermore, vocal tract configurations in shouted speech may have significant effects on speakers with unstable speech systems and related metrics of speech production variability, such as the lip-track (L-STI) or amplitude envelope (E-STI) based spatiotemporal indices (Howell *et al.*, 2009; Smith *et al.*, 1995). Such individual variants may contribute to the analysis of speaker-specific detail in shouted speech: for example, in forensic investigations), as shouted speech frequently occurs as evidence material in court cases (Blatchford and Foulkes, 2006).

Acknowledgments

This research was funded by the International Association for Forensic Phonetics and Acoustics: Research Grant "An investigation of the rhythmic acoustic differences between normal and shouted voices" (2015–2016).

Author Declarations

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

The study was conducted in line with the guidelines of the Ethics Committee of the Zürich University Faculty of Arts and Social Sciences.

Data Availability

The data that support the findings of this study are available from V.D. upon request.

References

Adel, Y., Nagel, S., Weissgerber, T., Baumann, U., and Macherey, O. (**2019**). "Pitch matching in cochlear implant users with single-sided deafness: effects of electrode position and acoustic stimulus type," Front. Neurosci. **13**, 1119.

Alghamdi, N., Maddock, S., Marxer, R., Barker, J., and Brown, G. J. (**2018**). "A corpus of audio-visual Lombard speech with frontal and profile views," J. Acoust. Soc. Am. **143**(6), EL523–EL529.

Arnoldner, C., Kaider, A., and Hamzavi, J. (**2006**). "The role of intensity upon pitch perception in cochlear implant recipients," Laryngoscope **116**(10), 1760–1765.

Baghel, S., Prasanna, S. R. M., and Guha, P. (**2021**). "Effect of high-energy voiced speech segments and speaker gender on shouted speech detection," in *2021 National Conference on Communications (NCC)*, Kanpur, India (IEEE, New York), pp. 1–6, available at https://ieeexplore.ieee.org/abstract/document/9530078?casa_token=cEAuN97wh_kAAAAA:_wVlwv5a4WUJ43hqSmBwPH4RK5g29306hIY6X828cp2QQeekiuCTGIDcDNWiEWif-RTqng5wIjM.

Baken, R. J., and Orlikoff, R. F. (**2000**). *Clinical Measurement of Speech and Voice* (Cengage Learning, Boston).

Baldner, E. F., Doll, E., and van Mersbergen, M. R. (**2015**). "A review of measures of vocal effort with a preliminary study on the establishment of a vocal effort measure," J. Voice **29**(5), 530–541.

Blatchford, H., and Foulkes, P. (**2006**). "Identification of voices in shouting," Int. J. Speech Lang. Law **13**(2), 241–254.

Boersma, P., and Weenick, D. (**2022**). "Praat: Doing phonetics by computer (version 6.2.23) [computer program]," http://www.praat.org (Last viewed October 8, 2022).

Bonnot, J.-F. P., and Chevrie-Muller, C. (**1991**). "Some effects of shouted and whispered conditions on temporal organization," J. Phon. **19**(3–4), 473–483.

Brockmann-Bauser, M., and Bohlender, J. E. (**2015**). (private communication).

Cushing, I. R., Li, F. F., Cox, T. J., Worrall, K., and Jackson, T. (**2011**). "Vocal effort levels in anechoic conditions," Appl. Acoust. **72**(9), 695–701.

Dromey, C., Ramig, L. O., and Johnson, A. B. (**1995**). "Phonatory and articulatory changes associated with increased vocal intensity in Parkinson disease: A case study," J. Speech. Lang. Hear. Res. **38**(4), 751–764.

Fischer, T., Schmid, C., Kompis, M., Mantokoudis, G., Caversaccio, M., and Wimmer, W. (**2021**). "Effects of temporal fine structure preservation on spatial hearing in bilateral cochlear implant users," J. Acoust. Soc. Am. **150**(2), 673–686.

Gangamohan, P., Gangashetty, S. V., and Yegnanarayana, B. (**2019**). "Subsegmental level analysis of high arousal speech using the zero-time windowing method," J. Acoust. Soc. Am. **145**(1), 551–561.

Howell, P., Anderson, A. J., Bartrip, J., and Bailey, E. (**2009**). "Comparison of acoustic and kinematic approaches to measuring utterance-level speech variability," J. Speech. Lang. Hear. Res. **52**(4), 1088–1096.

Jang, P.-S. (**2007**). "Designing acoustic and non-acoustic parameters of synthesized speech warnings to control perceived urgency," Int. J. Ind. Ergonom. **37**(3), 213–223.

Kobayashi, M., Hamada, Y., and Akagi, M. (**2022**). "Acoustic features correlated to perceived urgency in evacuation announcements," Speech Commun. **139**, 22–34.

Kong, Y.-Y., and Carlyon, R. P. (**2010**). "Temporal pitch perception at high rates in cochlear implants," J. Acoust. Soc. Am. **127**(5), 3114–3123.

Lagier, A., Legou, T., Galant, C., Bretèque, B. A. d. L., Meynadier, Y., and Giovanni, A. (**2017**). "The shouted voice: A pilot study of laryngeal physiology under extreme aerodynamic pressure," Logoped. Phoniatr. Vocol. **42**(4), 141–145.

Liénard, J.-S., and Benedetto, M.-G. D. (**1999**). "Effect of vocal effort on spectral properties of vowels," J. Acoust. Soc. Am. **106**(1), 411–422.

McKenna, V. S., and Stepp, C. E. (**2018**). "The relationship between acoustical and perceptual measures of vocal effort," J. Acoust. Soc. Am. **144**(3), 1643–1658.
Meister, H., Landwehr, M., Pyschny, V., Wagner, P., and Walger, M. (**2011**). "The perception of sentence stress in cochlear implant recipients," Ear Hear. **32**(4), 459–467.
Mittal, V. K., and Vuppala, A. K. (**2016a**). "Changes in shout features in automatically detected vowel regions," in *2016 International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India (IEEE, New York), pp. 1–5.
Mittal, V. K., and Vuppala, A. K. (**2016b**). "Significance of automatic detection of vowel regions for automatic shout detection in continuous speech," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Tianjin, China (IEEE, New York), pp. 1–5.
Mittal, V. K., and Yegnanarayana, B. (**2013**). "Effect of glottal dynamics in the production of shouted speech," J. Acoust. Soc. Am. **133**(5), 3050–3061.
Nawka, T., Verdonck-de Leeuw, I., De Bodt, M., Guimaraes, I., Holmberg, E., Rosen, C., Schindler, A., Woisard, V., Whurr, R., and Konerding, U. (**2009**). "Item reduction of the voice handicap index based on the original version and on European translations," Folia Phoniatr. Logop. **61**(1), 37–48.
Pohjalainen, J., Raitio, T., Yrttiaho, S., and Alku, P. (**2013**). "Detection of shouted speech in noise: Human and machine," J. Acoust. Soc. Am. **133**(4), 2377–2389.
Raitio, T., Suni, A., Pohjalainen, J., Airaksinen, M., Vainio, M., and Alku, P. (**2013**). "Analysis and synthesis of shouted speech," in *Proceedings of Interspeech 2013*, Lyon, France, pp. 1544–1548.
Rostolland, D. (**1982**). "Acoustic features of shouted voice," Acta Acust. Acust. **50**(2), 118–125, available at https://www.ingentaconnect.com/content/dav/aaua/1982/00000050/00000002/art00006.
Schulman, R. (**1989**). "Articulatory dynamics of loud and normal speech," J. Acoust. Soc. Am. **85**(1), 295–312.
Smith, A., Goffman, L., Zelaznik, H. N., Ying, G., and McGillem, C. (**1995**). "Spatiotemporal stability and patterning of speech movement sequences," Exp. Brain Res. **104**(3), 493–501.
Tak, S., and Yathiraj, A. (**2019**). "Comparison of intensity discrimination between children using cochlear implants and typically developing children," Int. Adv. Otol. **15**(3), 368–372.
Ternström, S., Bohman, M., and Södersten, M. (**2006**). "Loud speech over noise: Some spectral attributes, with gender differences," J. Acoust. Soc. Am. **119**(3), 1648–1665.
Traunmüller, H., and Eriksson, A. (**2000**). "Acoustic effects of variation in vocal effort by men, women, and children," J. Acoust. Soc. Am. **107**(6), 3438–3451.
Wagener, K., Brand, T., and Kollmeier, B. (**1999**). "Entwicklung und evaluation eines satztests für die deutsche sprache. i–iii. Design, optimierung und evaluation des Oldenburger satztests" ("Development and evaluation of a sentence test for the German language. i–iii. Design, optimization and evaluation of the Oldenburg sentence test"), Z. Audiologie (Audiological Acoust.) **38**, 4–15.
Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (**1991**). "Better speech recognition with cochlear implants," Nature **352**(6332), 236–238.
Winter, B. (**2013**). "Linear models and linear mixed effects models in R with linguistic applications," arXiv:1308.5499.
Wood, S. (**2017**). *Generalized Additive Models: An Introduction with R*, 2nd ed. (Chapman and Hall/CRC, Boca Raton, FL).
Wood, S. N. (**2011**). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," J. R. Stat. Soc. Ser. B: Stat. Methodol. **73**(1), 3–36.
Xue, Y., Marxen, M., Akagi, M., and Birkholz, P. (**2021**). "Acoustic and articulatory analysis and synthesis of shouted vowels," Comput. Speech Lang. **66**, 101156.
Zhang, C., and Hansen, J. H. L. (**2007**). "Analysis and classification of speech mode: Whispered through shouted," in *8th Annual Conference of the International Speech Communication Association, Interspeech 2007*, Antwerp, Belgium, pp. 2289–2292.