INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Voice quality as a turn-taking cue

Mattias Heldner[1], Marcin Włodarczak[1], Štefan Beňuš[2,3], Agustín Gravano[4,5]

[1]Department of Linguistics, Stockholm University, Sweden
[2]Constantine the Philosopher University in Nitra, Slovakia
[3]Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
[4]Departamento de Computación, FCEyN, Universidad de Buenos Aires (UBA), Argentina
[5]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina

[heldner,wlodarczak]@ling.su.se, sbenus@ukf.sk, gravano@dc.uba.ar

Abstract

This work revisits the idea that voice quality dynamics (VQ) contributes to conveying pragmatic distinctions, with two case studies to further test this idea. First, we explore VQ as a turn-taking cue, and then as a cue for distinguishing between different functions of affirmative cue words. We employ acoustic VQ measures claimed to be better suited for continuous speech than those in own previous work. Both cases indicate that the degree of periodicity (as measured by CPPS) is indeed relevant in the production of the different pragmatic functions. In particular, turn-yielding is characterized by lower periodicity, sometimes accompanied by presence of creaky voice. Periodicity also distinguishes between backchannels, agreements and acknowledgements.

Index Terms: voice quality, turn-taking cues, affirmative cue words

1. Introduction

This work revisits the idea that *voice quality dynamics* (VQ) contributes to conveying pragmatic distinctions (alongside other prosodic features), with two case studies to further test this idea. First, we explore VQ as a cue for different turn-taking transitions (i.e. smooth switches, holds, and backchannels [1, 2]), including the contribution of VQ to the recently proposed "go-signal" in turn-taking [3]. Then, we explore VQ as a cue for distinguishing between different functions of *affirmative cue words* (ACWs), and in particular between the frequent ACW categories *backchannel*, *acknowledgement* and *agreement* [4-6].

We follow Gobl, Ní Chasaide and colleagues [7, 8] in regarding short-term voice source variations, for instance between modal, breathy and creaky voice, as an additional dimension of prosodic expression together with more traditional prosodic features such as pitch and loudness dynamics.

The background for the first case study on cues for turn-continuations and turn-endings comprises a large body of work spanning at least the last half-century (see e.g. [1, 3] for recent reviews). In addition, a fair amount of work has been devoted to cues inviting backchannels (e.g. [2, 9, 10]). Many of these studies have in common that they investigate cues towards the very end of utterances. The investigated cues include prosodic ones, such as pitch movements and lengthening of segments, lexical, syntactic and pragmatic cues indicating completeness, gaze cues such as making and breaking eye contact, hand gestures, respiratory patterns, and so on. We note, however, that the studies exploring VQ as a turn-taking cue are relatively scarce. An early example is Ogden [11] who showed that creak has turn-yielding functions, and glottal stops have turn-holding functions in Finnish. Gravano and Hirschberg [2] investigated a large number of features as turn-taking cues in the Columbia Games Corpus in Standard American English. The features included contextual, intonation, as well as VQ features. With respect to VQ, they measured jitter and shimmer computed over voiced frames in the 500 ms interval preceding utterance ends, as well as noise-to-harmonics ratio (NHR) in the whole 500 ms interval. The VQ results showed that smooth switches were characterized by higher NHR, jitter and shimmer (all consistent with a lower degree of periodicity) than holds and backchannels. Brusco, et al. [1] replicated these findings also for Argentine Spanish. Similarly, Kane, et al. [12] explored various prosodic features, spectral similarity measures, as well as VQ features for distinguishing between pauses and gaps (turn-taking categories loosely related to the holds and smooth switches in [1, 2]). The VQ features described aspects of the glottal excitation and included the normalized amplitude quotient (NAQ) and the maxima dispersion quotient (MDQ). All features were extracted using methods available in the COVAREP repository [13]. However, contrary to the above-mentioned studies, it appears that the predictive power of these VQ features was marginal and also speaker dependent.

The rationale for the second case study on cues for ACWs, is that while backchannels, acknowledgements and agreements have different pragmatic functions (i.e. something like "Please continue" vs. "I believe what you say" vs. "I agree with what you say", respectively), they are typically expressed using the same lexical tokens, for example mm-hm, okay, uh-huh or yeah in American English, and mhm, no or uh-huh in Slovak; and sí, mm-hm or ok in Argentine Spanish [1, 4, 6]. These ACWs furthermore often occur in the same position in utterances—as single-word utterances. Thus, it seems plausible that speakers might employ other prosodic means, and possibly VQ to convey these distinctions. Gravano, et al. [4] explored a similar set of acoustic features as in [2] for distinguishing between different functions of affirmative cue words in the Columbia Games Corpus. Regarding VQ, their results showed that backchannels had a lower NHR than acknowledgements and agreements, and consequently that backchannels were overall less noisy and/or had less perturbed waveforms (and had a higher degree of periodicity). Notably, these findings are consistent both with backchannels being less breathy and/or being less creaky than the other ACWs.

Importantly, several of these previous studies used perturbation measures (i.e. jitter, shimmer, NHR) that are primarily

http://dx.doi.org/10.21437/Interspeech.2019-1592

intended for sustained vowels, and that are not recommended for running speech (cf. [14, 15]).

In this work, we re-analyzed the American English and Argentine Spanish material in [1, 2, 4] and added Slovak data [6]. We used VQ features that are more suitable for continuous speech. We did this to explore the idea that VQ contributes to conveying pragmatic distinctions, and to explore which pragmatic-prosody relationships hold cross-linguistically (or cross-culturally) and which might be language/culture specific.

2. Method

For this work, we analyzed data from three speech corpora, in three different languages. We used American English data from the Objects games in Columbia Games Corpus (e.g. [2, 4]), and in addition data from adaptations of the Objects games in Argentine Spanish [1], and in Slovak [6].

All three corpora include annotations of turn-taking transitions relevant for the first case study. Holds (H) are transitions where a *current speaker* continues after a short silence. Smooth switches (S) are transitions where a *next speaker* starts talking following a short silence. Backchannels (BC), finally, are transitions where a *next speaker* produces a short utterance with the primary function to invite the previous speaker to continue speaking. Note that transitions involving overlapping speech were excluded from analysis here. See Table 1 for the distribution of turn-taking transition types in the three corpora.

Table 1: *Distribution of turn-taking transition types (smooth switches S, holds H, backchannels BC) with valid VQ analyses in the three corpora.*

Language	**Turn-taking transition**	**N**	**%**
American English	S	1637	26.5
	H	4154	67.2
	BC	391	6.3
Argentine Spanish	S	2010	24.6
	H	5338	65.2
	BC	833	10.2
Slovak	S	1930	27.3
	H	4861	68.8
	BC	270	3.8

Table 2: *Distribution of affirmative cue word categories with valid VQ analyses, and a mapping between categories in the American English and Slovak data.*

American English	**N**	**%**	**Slovak**	**N**	**%**
Acknowledgement/ Agreement	1115	73.8	Acknowledge	99	22.5
			Acknowledge, but		
			Acknowledge, and start a new segment		
			Agreement	75	17.0
Backchannel	396	26.2	Backchannel	266	60.5

The American English and Slovak corpora also include annotations of ACWs relevant for the second case study, although with minor differences in the level of detail. In particular, while the Slovak corpus distinguishes between three different functions of acknowledgements plus one function for agreements, the American English corpus has a single Ack/Agr category for all of these acknowledgement and agreement functions (see Table 2 for a mapping of the ACW functions). In this work, we collapsed the different acknowledgement functions in the Slovak data into one category (Ack), but kept the agreement category (Agr) separate. The distribution of ACW categories in Table 2 includes only the three most frequent *(N > 20)* ACWs tokens "mmhm", "okay" and "yeah" in the American English data, and the most frequent "no" tokens only in the Slovak data.

We used acoustic VQ features suitable for continuous speech capturing the degree of periodicity in the speech signal Cepstral Peak Prominence Smooth (CPPS) [16, 17], and for identifying regions of creaky voice [18]. CPPS represents the amplitude of the most prominent cepstral peak (i.e. the dominant *rahmonic*) relative to the point with equal *quefrency* on the regression line through the smoothed *cepstrum* (in dB) [19]. We used Praat [20] and the method described in [19] to extract CPPS, and COVAREP [13] to extract regions of creaky voice. Both VQ features were extracted from the 500 ms preceding silences in the turn-taking transitions, and from the entire duration of the ACWs, and were described using medians. Note that the BC intervals in the two case studies represent different things: speech from another speaker preceding backchannels vs. the actual backchannels.

We used multinomial logistic regression to analyze the VQ features as turn-taking cues, and as cues for ACW function. We used Turn-taking transition with three levels (S, H, BC) as a categorical outcome variable in the first one. H was chosen as reference category as it does not involve speaker change. Furthermore, we used ACW function with two levels in the American English data (Ack/Agr, BC), and three levels in the Slovak data (Ack, Agr, BC) as outcome variable in the other one. Here BC was chosen as reference category. Median CPPS and median Creak in the relevant intervals were included as continuous predictors. The main effects, as well as their interaction were introduced one by one with a stepwise, forward entry procedure, given the criteria that they improve the fit of the model significantly.

3. Results

3.1. VQ as a turn-taking cue

The distributions of the VQ features (CPPS and Creak) preceding S, H, and BC turn-taking transition in the three languages are presented in Figure 1. The outcome of the corresponding multinomial logistic regression analyses is shown in Table 3.

We observed that the smooth switches generally had a lower degree of periodicity than the other transitions, both in terms of lower CPPS and in terms of higher Creak. Furthermore, the main effects of CPPS and Creak, as well as the interaction between CPPS and Creak were all included in the final model for American English (Table 3), although the Pseudo R-Squares indicated relatively low effect sizes. The odds ratios showed that an increase in CPPS decreased the odds for S relative to H, but increased the odds for BC relative to H. Creak as a main effect neither affected the odds for S relative to H, nor those for BC relative to H in the model. However, the significant interaction between the predictors showed that an increase in Creak increased the odds for S relative to H, but not for BC relative to H.

For Argentine Spanish, only the main effects were included in the final model. Again, the effect sizes were low.

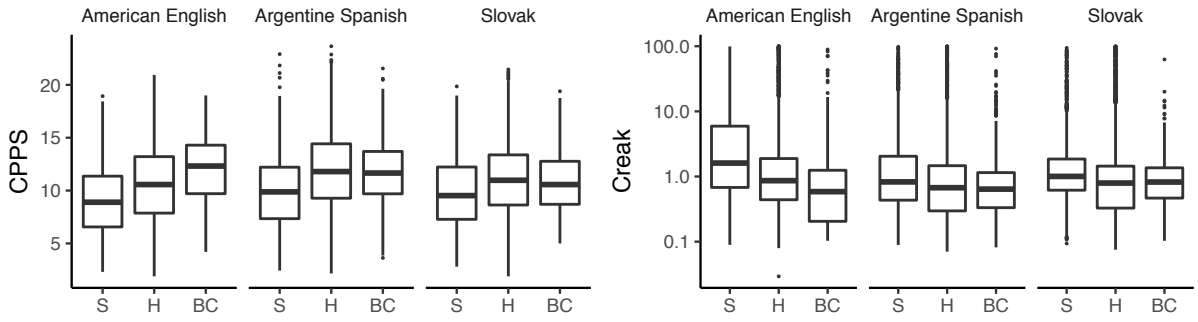

Figure 1: *Boxplots of median CPPS (left panel) and median Creak (right panel) in the 500 ms preceding silences in smooth switch (S), hold (H), and backchannel (BC) turn-taking transitions in American English, Argentine Spanish, and Slovak.*

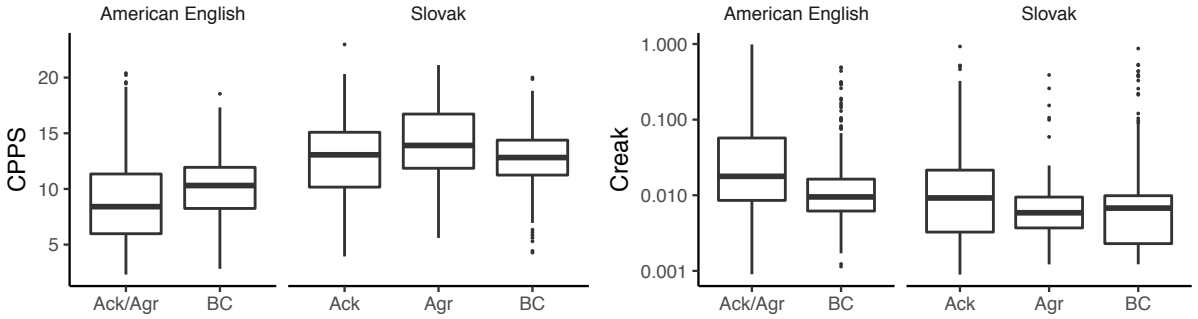

Figure 2: *Boxplots of median CPPS (left panel) and median Creak (right panel) in acknowledgement/agreement (Ack/agr) and backchannel (BC) affirmative cue words in American English, and acknowledgement (Ack) agreement (Agr) and backchannel (BC) affirmative cue words in Slovak.*

Table 3: *Multinomial logistic regression with turn-taking transition type as dependent variable (H as reference category), and CPPS and Creak as covariates, for American English, Argentine Spanish and Slovak. * p < .05, ** p < .01, *** p < .001.*

				95% CI for Odds Ratio		
Language	**Turn-taking transition**		**B (SE)**	**Lower**	**Odds Ratio**	**Upper**
American English	S vs. H	Intercept	.13 (.10)			
		CPPS	-.11 (.01) ***	.88	.89	.91
		Creak	-.92 (.47)	.16	.40	1.00
		Creak * CPPS	.25 (.06) ***	1.13	1.28	1.45
	BC vs. H	Intercept	-3.49 (.21) ***			
		CPPS	.10 (.02) ***	1.07	1.11	1.14
		Creak	-1.88 (1.76)	.01	.15	4.78
		Creak * CPPS	.11 (.20)	.75	1.12	1.66
Note: R^2 = .05 (Cox & Snell), .06 (Nagelkerke). Model $\chi^2(6) = 330.49$, $p < .001$.						
Argentine Spanish	S vs. H	Intercept	.82 (.09) ***			
		CPPS	-.16 (.01) ***	.84	.85	.86
		Creak	-1.92 (.28) ***	.09	.15	.25
	BC vs. H	Intercept	-1.56 (.14) ***			
		CPPS	-.02 (.01)	.96	.98	1.00
		Creak	-3.34 (.75) ***	.01	.03	.15
Note: R^2 = .06 (Cox & Snell), .07 (Nagelkerke). Model $\chi^2(4) = 471.48$, $p < .001$.						
Slovak	S vs. H	Intercept	.39 (.09) ***			
		CPPS	-.12 (.01) ***	.87	.88	.90
		Creak	-1.23 (.28) ***	.17	.29	.50
	BC vs. H	Intercept	-2.37 (.23) ***			
		CPPS	-.04 (.02) *	.93	.96	.99
		Creak	-3.76 (1.40) **	.00	.02	.36
Note: R^2 = .03 (Cox & Snell), .04 (Nagelkerke). Model $\chi^2(4) = 224.38$, $p < .001$.						

Table 4: *Multinomial logistic regression with affirmative cue word function as dependent variable (BC as reference category), and CPPS and Creak as covariates, for American English and Slovak. * $p < .05$, ** $p < .01$, *** $p < .001$.*

Language	ACW function		B (SE)	95% CI for Odds Ratios		
				Lower	Odds Ratios	Upper
American English	Ack/Agr vs. BC	Intercept	1.37 (.20) ***			
		CPPS	-.06 (.02) ***	.91	.94	.98
		Creak	6.18 (1.27) ***	40.47	483.48	5775.82
Note: R^2 = .06 (Cox & Snell), .09 (Nagelkerke). Model $\chi^2(2) = 92.81$, $p < .001$.						
Slovak	Ack vs. BC	Intercept	-1.02 (.49) *			
		CPPS	.00 (.04)	.93	1.00	1.08
	Agr vs. BC	Intercept	-3.16 (.61) ***			
		CPPS	.14 (.04) **	1.06	1.15	1.25
Note: R^2 = .03 (Cox & Snell), .03 (Nagelkerke). Model $\chi^2(2) = 11.67$, $p < .01$.						

As in the American English model, an increase in CPPS lowered the odds for S relative to H, but here it did not affect the odds for BC relative to H. Contrary to the American English model, an increase in Creak decreased the odds of S relative to H as well as the odds of BC relative to H.

For Slovak, only the main effects were included in the final model. As in the other two languages, an increase in CPPS decreased the odds for S relative to H. Similar to the Argentine Spanish model but contrary to the American English model, an increase in Creak decreased the odds of S relative to H as well as the odds of BC relative to H.

3.2. VQ as an ACW function cue

The distributions of the VQ features (CPPS and Creak) in the different ACW functions in American English and Slovak are presented in Figure 2. The outcome of the corresponding multinomial logistic regression analyses is shown in Table 4. First, we observed that the American English ACWs generally had a lower degree of periodicity than the Slovak ones, both in terms of lower CPPS and in terms of higher Creak. We also observed lower CPPS as well as higher Creak values in Ack/Agr than in BC in the American English data. Furthermore, only the main effects were included in the final model for American English, and an increase in CPPS decreased the odds for Ack/Agr, whereas an increase in Creak increased the odds for Ack/Agr (relative to BC). The analyses of the Slovak data showed that only the main effect of CPPS was included in the final model. Contrary to the American English data, an increase in CPPS increased the odds for Agr, whereas CPPS did not affect the odds for Ack.

4. Discussion

The results from all three languages indicate that lower CPPS, and hence a lower periodicity, before silence is associated with smooth switches. VQ in terms of a lower periodicity is thus potentially a cross-linguistic turn-yielding cue and/or contributes to "go-signals" in turn-taking [3]. Alternatively, these results can of course also be interpreted as an association between higher periodicity before silence and turn-holds–that is as a "stop signal" in turn-taking. The results regarding VQ as a backchannel inviting cue, however, are weak and conflicting across languages.

The observations of lower periodicity in smooth switches compared to holds confirms earlier findings of a higher Noise-To-Harmonics ratio, higher jitter and higher shimmer in the same comparison based on parts of the data used here [1]. They are also consistent with the results of creak as a turn-yielding cues for Finnish [11].

We note that lower periodicity (as indicated by a higher NHR or a lower CPPS) could be the effect of fundamental frequency or amplitude perturbations (e.g. captured by jitter or shimmer measures), but it could also result from presence of noise, for example in a breathy voice quality. Importantly, the American English data shows that the lower CPPS values were accompanied by higher creak, which also results in lower periodicity. We note that although the CPPS and Creak measures used here are correlated (R^2 range between .05 and .13 in our data), there is independent variation as well. While high creak is generally associated with low CPPS, there is substantial CPPS variation when Creak is low, as well as substantial Creak variation when CPPS is low. Thus, low CPPS combined with low Creak could, for instance, be an indication of breathy voice. In any case, acoustic measures of Creak provide additional information for the interpretation of periodicity.

The results from the two corpora including ACW annotations indicate that VQ is potentially a cue of ACW function. However, it appears that American English and Slovak differ in how VQ contributes to conveying ACW functions. Acknowledgements and agreements had lower CPPS and higher Creak values, and thus lower periodicity than backchannels in American English. For Slovak, the results were reversed with lower CPPS values (and hence lower periodicity) in the backchannels, and Creak did not contribute to cueing the ACW functions. Thus, it appears that this use of VQ for cueing pragmatic distinctions may be language and/or culture specific. Furthermore, the ACW results for American English here replicate the previous finding that acknowledgements and agreements in this corpus had higher NHR, and consequently lower periodicity than backchannels [4]. But we consider it an important contribution of the present study to show that the lower periodicity could to some extent be explained by presence of Creak. We conclude that the findings from our two case studies support the idea that voice prosody contributes to conveying pragmatic function, and note that presence of creaky voice must be considered when evaluating periodicity measures. From a practical viewpoint, given that CPPS and creaky voice can now be directly measured with tools such as COVAREP, our findings may favor these features over more traditional ones such as jitter, shimmmer or NHR for turn-taking applications.

5. Acknowledgements

This work was partly funded by the project MAW project 2017.0034 to the first author, a Christian Benoît Award to the second author, and a VEGA 2/0161/18 grant to the third author. This material is based upon work supported by CONICET, ANPCYT PICT 2014-1561, and the Air Force Office of Scientific Research under award no. FA9550-18-1-0026.

6. References

[1] Brusco, P., Pérez, J. M., and Gravano, A., "Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish," in *Proceedings Interspeech 2017*, Stockholm, Sweden, 2017, pp. 2351–2355.

[2] Gravano, A. and Hirschberg, J., "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, pp. 601–634, 2011.

[3] Barthel, M., Meyer, A. S., and Levinson, S. C., "Next speakers plan their turn early and speak after turn-final "go-signals"," *Frontiers in Psychology*, vol. 8, p. 393, 2017.

[4] Gravano, A., Hirschberg, J., and Beňuš, Š., "Affirmative cue words in task-oriented dialogue," *Computational Linguistics*, vol. 38, pp. 1–39, 2012.

[5] Beňuš, Š., Gravano, A., and Hirschberg, J., "The prosody of backchannels in American English," in *Proceedings ICPhS 2007*, Saarbrücken, Germany, 2007, pp. 1065–1068.

[6] Beňuš, Š., "The prosody of backchannels in Slovak," in *Proceedings Speech Prosody 2016*, Boston, USA, 2016, pp. 415–419.

[7] Gobl, C. and Ní Chasaide, A., "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle*, et al.*, Eds. Oxford: Wiley-Blackwell, 2012, pp. 378–423.

[8] Gobl, C., Yanushevskaya, I., and Chasaide, A. N., "The relationship between voice source parameters and the maxima dispersion quotient (MDQ)," in *Proceedings Interspeech 2015*, Dresden, Germany, 2015, pp. 2337–2341.

[9] Ward, N. G., *Prosodic Patterns in English Conversation*. Cambridge, UK: Cambrige University Press, 2019.

[10] Ward, N. and Tsukahara, W., "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, pp. 1177-1207, 2000.

[11] Ogden, R., "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association*, vol. 31, 2002.

[12] Kane, J., Yanushevskaya, I., de Looze, C. l., Vaughan, B., and Ní Chasaide, A., "Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions," in *Proceedings Interspeech 2014*, Singapore, 2014, pp. 333–337.

[13] Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S., "COVAREP - A collaborative voice analysis repository for speech technologies," in *Proceedings ICASSP 2014*, Florence, Italy, 2014, pp. 960–964.

[14] Sprecher, A., Olszewski, A., Jiang, J. J., and Zhang, Y., "Updating signal typing in voice: addition of type 4 signals," *Journal of the Acoustical Society of America*, vol. 127, pp. 3710-3716, Jun 2010.

[15] Titze, I. R. (1995). *Workshop on Acoustic Voice Analysis: Summary Statement*. Available: http://www.ncvs.org/freebooks/summary-statement.pdf

[16] Hillenbrand, J. and Houde, R. A., "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech Language and Hearing Research*, vol. 39, 1996.

[17] Ferrer Riesgo, C. A. and Nöth, E., "What makes the Cepstral Peak Prominence different to other acoustic correlates of vocal quality? [Article in Press]," *Journal of Voice*, Jan 22 2019.

[18] Drugman, T., Kane, J., and Gobl, C., "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech & Language*, vol. 28, pp. 1233–1253, 2014.

[19] Watts, C. R., Awan, S. N., and Maryn, Y., "A comparison of cepstral peak prominence measures from two acoustic analysis programs," *Journal of Voice*, vol. 31, pp. 387 e1–387 e10, May 2017.

[20] Boersma, P. and Weenink, D. (2019), "Praat: doing phonetics by computer [Computer program]" (Version 6.0.39). Available: http://www.praat.org/