VOCAL EFFORT IN HUMAN INTERACTIONS WITH VOICE-AI

Leah Bradshaw[1], Valeriia Perepelytsia[1] and Volker Dellwo[1]
[1]Department of Computational Linguistics, University of Zurich, Switzerland
leah.bradshaw@uzh.ch

ABSTRACT

Speech adaptations occur frequently in the presence of perceived communication barriers. Modern technological advancements have brought with them new interlocutors for human speakers with the introduction of voice-AI assistants. Findings have shown that voice-AI-directed speech is characterised by an increase in vocal effort resulting from the presumed capabilities of these systems for understanding speech. However, studies focus solely on voice-AI assistants which perform speech recognition. In this study, we present an acoustic analysis of speaker interactions with two voice-AI systems with different goals (speech interpretation vs. speaker verification). Using f_0 mean and range as acoustic correlates of vocal effort, we found that speakers show some evidence of increased vocal effort towards voice-AI systems regardless of final task, however, this is enhanced by speech intelligibility goals. This finding is interpreted to suggest that voice-AI-directed speech globally exhibits increased vocal effort, but task plays a clear role in the extent of this.
(149 words)

Keywords: speech communication, vocal effort, voice-AI, speaker accommodation

1. INTRODUCTION

Speech in challenging conditions, i.e., in noisy environments or towards less-proficient interlocutors, is frequently observed to be produced with increased vocal effort, characterised by increased fundamental frequency (f_0) mean and range [e.g., 1-7]. The introduction of voice activated artificially intelligent, or voice-AI, assistants has created a new kind of challenging interlocutor, which speakers assume to require more effortful talk [7–11]. However, human interactions with these systems have thus far only been explored using speech recognition technologies. Contrastively, voice-AI systems can perform numerous tasks which do not process speech in the same way, i.e., speaker verification. It remains to be explored how speakers interact more globally with voice-AI systems and how the specific task plays a role in the amount of vocal effort employed.

In this study, we investigated if speakers employ differing vocal effort towards voice-AI systems according to task. We designed a Wizard of Oz experiment [e.g., 12–14] in which participants interacted with two different mock voice-AI systems; a Speech Recogniser, which aimed to understand their speech; and a Speaker Recogniser, which aimed to identify them from their voice. We present here the findings of an acoustic analysis which explores two acoustic correlates of vocal effort (f_0 mean and range), to examine differences in how speakers interact with the two systems.

1.1. Vocal effort in speech communication

Increased vocal effort occurs frequently in challenging communicative environments, for example, when speakers shift from a neutral speech to shouted speech in the presence of noise [e.g., 1–5] or when the distance between interlocutors is increased [e.g., 6, 15]. One common acoustic correlate associated with this phenomenon in the aforementioned studies, is increased f_0 mean and range. Vocal adaptations corresponding to increases in f_0 mean and range also occur in the presence of challenging interlocutors or to enhance speech intelligibility, i.e., towards children [16–17] or voice-AI systems [7], and are attributed in part to increased vocal effort. However, in these instances speech intelligibility requirements are confounded with challenging interlocutors, thus the influence of each individual factor on vocal effort cannot be isolated.

1.2. Voice-AI-directed speech

Studies have shown numerous interlocutors require speech adaptations for efficient communication. Children, hearing-impaired listeners, and, more recently, machines all pose intelligibility challenges for speakers [16–21]. The introduction of the voice-AI assistant has commanded a new challenging machine interlocutor for humans, which invokes a speech style that is louder [22–23] and contains different vowel formant characteristics [23] compared to human-directed speech. Further, findings have shown increases in f_0 mean and range in voice-AI-directed speech [22] and have attributed these in part to an increase in vocal effort in these interactions [7].

Thus far, voice-AI-directed speech has solely considered interactions with speech recognition technologies, such as Amazon's Alexa or Apple's

Siri, which require interpretation of speech input. Therefore, it is plausible that the previously observed voice-AI-directed speech characteristics are a by-product of the systems requirement for speech intelligibility. However, in fields such as banking, there is an increasing use of voice-AI systems for voice identification/verification purposes, a task which considers the quality of the voice rather than the content of the speech. As such, the question arises whether speakers would employ the same vocal adaptations towards these kinds of voice-AI systems which offer a similarly challenging interlocutor, but without the necessity for speech intelligibility.

1.3. Research hypotheses

This study investigates vocal effort in interactions with two different voice-AI systems, a Speech Recogniser, and a Speaker Recogniser. Previous findings show speakers employ greater vocal effort, measured by an increase in f_0 mean and range, in interactions with speech recogniser voice-AI systems. However, the extent to which this vocal effort relates to the task of speech understanding or to the generally challenging interlocutor remains unknown. We predict that, if vocal effort is employed in voice-AI interactions due to the perceived communicative barriers of the system, vocal effort will be similar in both interactions. However, if speech intelligibility challenges necessitate enhanced vocal effort, we expect lesser vocal effort will be employed towards the Speaker Recogniser.

2. METHODS

2.1. Data collection

39 Swiss German speakers (20 F) completed a Wizard of Oz experiment, an expansion on the experiment design introduced in [24], designed using Gorilla Experiment Builder [25]. Participants spoke to two mock voice-AI systems; a mock speaker recognition system (SpkrRec) with a male text-to-speech voice, Verifico, which aimed to recognise the individual from their voice input; and a mock speech recognition system (SpchRec) with a female text-to-speech voice, Vicky, whose task was to interpret their speech. Participants spoke 34 different prompts to each of the systems which then offered either an immediate correct response, or a misrecognition. Each trial was repeated until the "system" was successful. Responses were assigned randomly by the experiment software, but participants would only ever receive a maximum of misrecognitions per prompt. All speakers spoke Swiss Standard German throughout the experiment [26]. Full details of the experimental procedure can be found in [27].

Prior to the interactions, speakers were recorded reading a list of the same 34 sentences used as prompts in the interaction tasks which were used as baseline productions for each speaker.

2.2. Data analysis

Analysis was completed using only first productions of a sentence, resulting in 102 sentences per speaker. Repetitions of sentences following an error response were excluded so that all sentences were produced following positive feedback (a recognition) or no feedback in the case of the read speech condition.

Acoustic measurements were automatically extracted over each target sentence in each of the three tasks using a Praat script [28]. Fundamental frequency (f_0) minimum and maximum values were extracted for each sentence in logHz, to calculate the f_0 range ($f_{0max} - f_{0min}$). Mean f_0 was calculated at 15 equidistant intervals across each sentence, also in logHz, and averaged to calculate the mean f_0 value. All measures were taken using gender-specific pitch ranges (50-200Hz for males; 75-400Hz for females). All measures were also z-scored prior to statistical analysis to account for individual speaker differences.

2.3. Statistics

Each acoustic measure was subjected to statistical analysis to explore the differences between interactions with each voice-AI system. Mean f_0 and f_0 range were modelled in separate linear mixed effects models with the *lme4* R package [29], with identical model structure: fixed effects of *Task* (SpkrRec, SpchRec, Read), *Gender* (Male, Female), plus by-Sentence and by-Speaker random intercepts.

3. RESULTS

3.1. Mean f_0

Figure 1 contains the distribution of mean f_0 for speakers' productions of each sentence across the three tasks. Model output showed a statistically significant increase in mean f_0 in both tasks, compared to the Reading task (SpchRec: $\beta = 0.553$, SpkrRec: $\beta = 0.1738$, both $p < 0.0001$). The main effect of Gender failed to reach significance. Post-hoc tests to assess the directionality of this relationship were conducted in the form of pairwise t-test comparisons with Bonferroni correction for multiple hypothesis testing. Findings confirmed a statistically significant difference between f_0 mean in the SpchRec and SpkrRec tasks compared to the Reading task, as well as a statistically significant difference between the two voice-AI tasks (all $p < 0.001$).

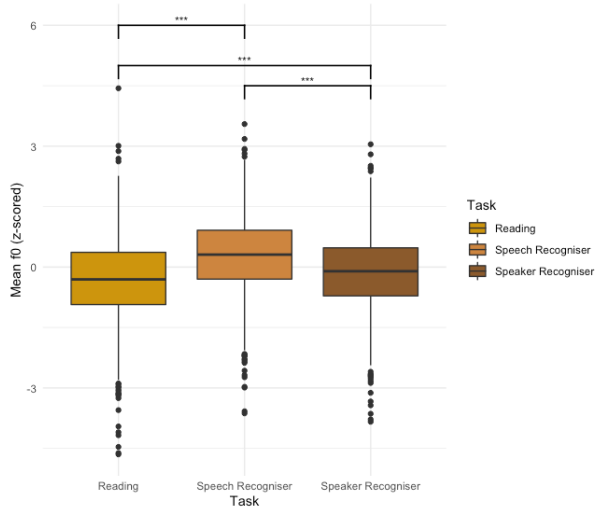

Figure 1: Boxplots showing the median, range, and inter-quartile range of the distribution of scaled mean f_0 values for speakers' production of each utterance by task.

3.2. f_0 range

Figure 2 shows the distribution of f_0 range for speakers' productions of each sentence in each of the three tasks. The model output showed a statistically significant increase in f_0 range in the SpchRec task, compared to the Reading task ($\beta = 0.0895$, $p = 0.002$). However, the difference between the SpkrRec and Reading tasks was not statistically significant. The main effect of Gender also failed to reach significance. Post-hoc analysis of f_0 range by Task was again completed using pairwise t-test comparisons with Bonferroni correction. The output showed a statistically significant difference between the SpchRec and the Reading task ($p < 0.001$), but not between the Reading and SpkrRec tasks ($p = 0.5396$), or the SpchRec and SpkrRec tasks ($p = 0.2573$).

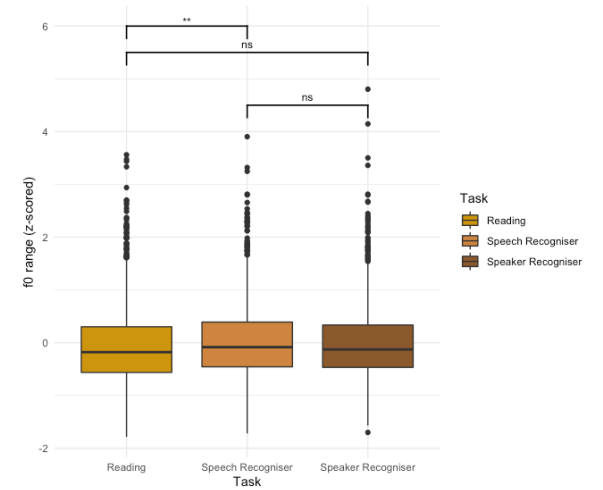

Figure 2: Boxplots showing the median, range, and inter-quartile range for the distribution of scaled f_0 range values for speakers' produced utterances by task.

3.3. Accommodation to voice-AI

We also tested for a potential confound in the finding that mean f_0 was substantially higher in the SpchRec task. Given the voice in this task (Vicky) was a female voice, compared to the male voice (Verifico) in the SpkrRec task, speaker accommodation towards the voice-AI system would increase in average speaker mean f_0, which may account for this finding.

To explore this, we calculated two distances. The distance between the speaker's averaged mean f_0 and the mean f_0 of each voice-AI system *prior* to and *during* the interaction. Mean f_0 for each voice-AI was calculated in the same way as previously stated, and speaker values were averaged across each task. The measures were as follows:

Distance 1: Euclidean distance between the mean f_0 for Vicky/Verifico and each speakers' pre-interaction productions (Reading task) (voice-AI mean f_0 – participants' mean $f_{0\,[Read]}$)

Distance 2: Euclidean distance between the mean f_0 for Vicky/Verifico and each speakers' production in the corresponding interaction task (Vicky mean f_0 – participants' mean $f_{0\,[SpchRec]}$ | Verifico mean f_0 – participants' mean $f_{0\,[SpkrRec]}$)

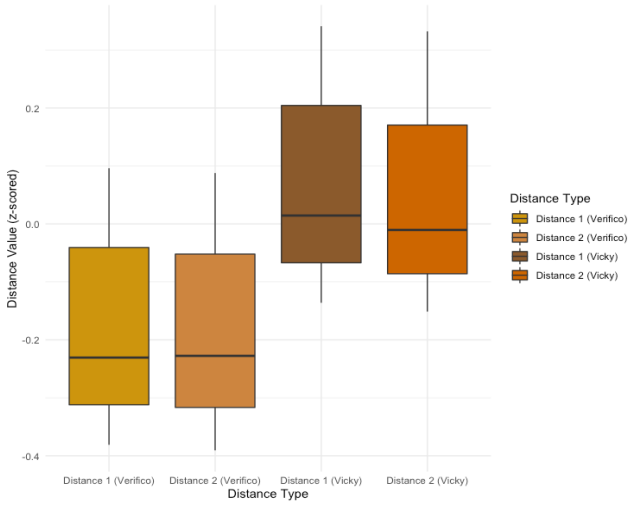

Figure 3: Boxplots showing medians, ranges, and inter-quartile ranges for scaled distances between averaged mean f_0 for each speaker and mean f_0 of the voice-AI systems, both prior and during the interaction.

Figure 3 shows the comparison between *Distance 1* and *Distance 2* for each voice-AI system. We ran a two-way repeated measures ANOVA to test for an interaction between *Time* (e.g., prior or during the interaction with the voice-AI) and *System* (e.g., Vicky or Verifico). The outcome was statistically significant ($p = 0.0067$), suggesting both factors influenced the difference between the distances. To assess accommodation towards each voice-AI system, we conducted further pairwise t-test with Bonferroni correction for repeated measures. Findings showed

that the difference between *Distance 1* and *2* was significant for both voice-AI systems (Vicky: p = 0.000196, Verifico: p = 0.032).

4. DISCUSSION

Overall, we find evidence to suggest that increased vocal effort, measured using f_0 mean and range, is a general characteristic of voice-AI-directed speech. However, we observe differences in the extent of this depending on the task. Namely, we observed a significant increase in both f_0 mean and range in the SpchRec task, but only an increase in f_0 mean in the SpkrRec task. Further, the increase in f_0 mean in the SpchRec task was significantly larger than that observed in the SpkrRec task. Therefore, it is plausible that, while human interactions with voice-AI systems can be generally associated with enhancements in vocal effort, the necessity for speech intelligibility further enhances the amount of vocal effort in speech.

The lesser vocal effort observed in the interactions with the SpkrRec system could be attributed to the lack of requirement for speech intelligibility. Previous findings of increased vocal effort towards challenging interlocutors are confounded with speech intelligibility challenges [e.g., 7, 16–17]. For instance, increased vocal effort observed in speech towards children could arise from the fact that the interlocutor is less proficient, or the desire for speech interpretation. Therefore, the vocal effort exhibited towards the SpkrRec could represent that induced solely by the presence of a challenging interlocutor.

Contrastively, it may be that speakers did not deem the SpkrRec to be as challenging an interlocutor as the SpchRec, due to a lack of prior experience with the system. From post-experiment discussions, it was clear that participants were already familiar with Speech Recognition systems, but not Speaker Recognition systems. Therefore, the greater vocal effort could correspond to prior assumptions that the SpchRec system would be a challenging interlocutor, compared to a lack of prior assumptions for the SpkrRec.

A potential confounding with these findings is the comparison with the monologue speech in the Reading task. Although it is likely that a voice-AI interlocutor would compel more effortful speech, it is possible that this slight increase in vocal effort we observe is due to the presence of the interlocutor in the SpkrRec task. Further research is necessary to explore to the effect of interlocutor presence or absence.

Further consideration was given to possible speaker accommodation towards the voice-AI system which may have commanded the higher f_0 mean in the SpchRec task. However, we observed significant convergence towards both systems and thus cannot solely attribute the increased mean f_0 to speaker accommodation. Further, it is equally not unexpected that participants would converge towards the systems. Previous findings show speaker accommodation is a frequently employed technique for enhancing intelligibility [e.g., 30-31], and speakers have previously exhibited phonetic and prosodic alignment towards machine interlocutors [e.g., 32-34].

Finally, we note that f_0 mean and range only offer an introductory look at vocal effort differences in these two speech styles. Additional research using a more comprehensive set of measures, including for instance, speech intensity/loudness, is necessary to fully examine these differences.

Overall, the above represents a multitude of plausible interpretations for these findings given previous ideas of speech communication and vocal effort. We observed clear differences in speaker f_0 mean and range in interactions with two different kinds of voice-AI systems, however, which factor constitutes to this finding is unclear. Nonetheless, it is conceivable that speakers globally adopt more effortful talk when speaking to voice-AI systems, which is enhanced by speech intelligibility requirements.

5. CONCLUSION

The present study investigated vocal effort in interactions with two voice-AI systems which perform different tasks. Findings showed a tendency for increased vocal effort, captured using f_0 mean and range, in interactions with both of the voice-AI systems compared to the Reading task. However, substantially greater vocal effort was employed when speech intelligibility was required. These findings are interpreted to suggest that voice-AI-directed speech globally exhibits increased vocal effort, but the extent to which this occurs can differ depending on the task. Future research should consider further factors influencing vocal effort, namely, the goal of the interaction and the presence of an interlocutor. Further discussions about the different kinds of voice-AI systems and human interactions with them, would also be beneficial.

6. ACKNOWLEDGMENTS

This research was supported by the Swiss National Science Foundation in the project #185399 “The dynamics of indexical information in speech and its role in speech communication and speaker recognition.”

7. REFERENCES

[1] M. Jessen, O. Koster, and S. Gfroerer, "Influence of vocal effort on average and variability of fundamental frequency," *IJSLL*, vol. 12, no. 2, Art. no. 2, Aug. 2005.

[2] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *JASA*, vol. 84, no. 3, pp. 917–928, Sep. 1988.

[3] Z. S. Bond, T. J. Moore, and B. Gable, "Acoustic–phonetic characteristics of speech produced in noise and while wearing an oxygen mask," *JASA*, vol. 85, no. 2, pp. 907–912, Feb. 1989.

[4] J. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *JASA*, vol. 93, no. 1, pp. 510–524, Jan. 1993.

[5] S. Koster, "Acoustic-phonetic aspects of Lombard Speech for different text styles," *The Phonetician*, vol. 85, pp. 9–16, 2002.

[6] P. Bottalico, "Speech Adjustments for Room Acoustics and Their Effects on Vocal Effort," *Journal of Voice*, vol. 31, no. 3, p. 392.e1-392.e12, May 2017.

[7] M. Cohn and G. Zellou, "Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments," *Frontiers in Communication*, vol. 6, 2021.

[8] H. H. Clark and G. L. Murphy, "Audience Design in Meaning and Reference," in *Advances in Psychology*, vol. 9, J.-F. Le Ny and W. Kintsch, Eds. North-Holland, 1982, pp. 287–299.

[9] S. Oviatt, G.-A. Levow, E. Moreton, and M. MacEachern, "Modeling global and focal hyperarticulation during human–computer error resolution," *JASA*, vol. 104, no. 5, pp. 3080–3098, Nov. 1998.

[10] H. P. Branigan, M. J. Pickering, J. Pearson, J. F. McLean, and A. Brown, "The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers," *Cognition*, vol. 121, no. 1, pp. 41–57, Oct. 2011.

[11] B. R. Cowan, H. P. Branigan, M. Obregón, E. Bugis, and R. Beale, "Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human−computer dialogue," *Int. J. Hum. Comput. Stud.* vol. 83, pp. 27–42, Nov. 2015.

[12] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of Oz studies — why and how," *Knowledge-Based Systems*, vol. 6, no. 4, pp. 258–266, Dec. 1993.

[13] D. DeVault, J. Mell, and J. Gratch, "Toward Natural Turn-Taking in a Virtual Human Negotiation Agent," in *2015 AAAI Spring Symposium Series*, Mar. 2015.

[14] S. Oviatt, C. Darves, and R. Coulston, "Toward adaptive conversational interfaces: Modeling speech convergence with animated personas," *ACM Trans. Comput.-Hum. Interact.*, vol. 11, no. 3, pp. 300–328, Sep. 2004.

[15] J.-S. Liénard and M.-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *JASA*, vol. 106, no. 1, pp. 411–422, Jul. 1999.

[16] A. R. Bradlow, N. Kraus, and E. Hayes, "Speaking Clearly for Children With Learning Disabilities," *JSLHR*, vol. 46, no. 1, pp. 80–97, Feb. 2003.

[17] M. Uther, M. A. Knoll, and D. Burnham, "Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech," *Speech Commun.*, vol. 49, no. 1, pp. 2–7, Jan. 2007.

[18] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking Clearly for the Hard of Hearing II," *JSLHR*, vol. 29, no. 4, pp. 434–446, Dec. 1986.

[19] R. Lunsford, S. Oviatt, and A. M. Arthur, "Toward open-microphone engagement for multiparty interactions," in *Proceedings of the 8th international conference on Multimodal interfaces - ICMI '06*, Banff, Alberta, Canada, 2006, p. 273.

[20] D. Burnham, S. Joeffry, and L. Rice, "Computer-and human-directed speech before and after correction," *space*, vol. 6, p. 7, 2010.

[21] C. Mayo, V. Aubanel, and M. Cooke, "Effect of prosodic changes on speech intelligibility," in *Proceedings of INTERSPEECH*, Portland Oregon USA, Sept. 2012.

[22] E. Raveh, I. Steiner, I. Siegert, I. Gessinger, and B. Möbius, "Comparing phonetic changes in computer-directed and human-directed speech," *Elektronische Sprachsignalverarbeitung 2019*, pp. 42–49, 2019.

[23] I. Siegert and J. Krüger, "'Speech Melody and Speech Content Didn't Fit Together'—Differences in Speech Behavior for Device Directed and Human Directed Interactions," in *Advances in Data Science: Methodologies and Applications*, Springer, 2021, pp. 65–95.

[24] V. Dellwo, E. Pellegrino, L. He, and T. Kathiresan, 'The dynamics of indexical information in speech: Can recognizability be controlled by the speaker?', *AUC Philol.*, vol. 2019, no. 2, pp. 57–75, Oct. 2019.

[25] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, 'Gorilla in our midst: An online behavioral experiment builder', *Behav. Res. Methods*, vol. 52, no. 1, pp. 388–407, Feb. 2020.

[26] I. Hove, *Die Aussprache der Standardsprache in der deutschen Schweiz*. DE GRUYTER, 2002.

[27] V. Perepelytsia, L. Bradshaw and V. Dellwo, "IDEAR: A speech database of identity-marked, clear and read speech,", *Proceedings of ICPhS*, 2023.

[28] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer'. 2022. [Online].

[29] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," Journal of Statistical Software, 67(1), 1-48, 2015.

[30] R. L. Street and H. Giles, "Speech accommodation theory: A social cognitive approach to language and speech behavior," *Social cognition and communication*, vol. 193226, pp. 193–226, 1982.

[31] C. Gallois, T. Ogay, and H. Giles, "Communication Accommodation Theory," *Communication Accommodation Theory*, p. 28.

[32] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *Proceedings of ICPHS*, 2003, vol. 3, pp. 833–836.

[33] N. Suzuki and Y. Katagiri, "Prosodic alignment in human–computer interaction," *Connection Science*, vol. 19, no. 2, pp. 131–141, Jun. 2007.

[34] I. Gessinger, B. Möbius, N. Fakhar, E. Raveh, and I. Steiner, 'A Wizard-of-Oz experiment to study phonetic accommodation in human-computer interaction', presented at *ICPhS*, Melbourne, 2019, pp. 1475–1479.