forum acusticum 2026
8-12 September · Graz, Austria

CHARISMA IN A FEW SECONDS: PROSODIC PROMINENCE AND THE PERCEPTION OF VOCAL CHARISMA IN SHORT CONVERSATIONAL SPEECH

Julian Linke*[1], Oliver Niebuhr*[2], Martin Wellenhof-Karner[1], Barbara Schuppler[1]

[1] *Signal Processing and Speech Communication Laboratory, Graz University of Technology*
[2] *Centre for Industrial Electronics, University of Southern Denmark*

Abstract

This study examines whether vocal charisma can be assessed from ultra-short conversational speech and whether prosodic prominence contributes to such judgments. Thirty listeners rated charisma, likeability, self-confidence, and passion in 92 spontaneous Austrian German excerpts lasting 1–2 s. Charisma ratings showed a coherent perceptual structure and were most closely associated with likeability. They also correlated strongly with PICSA-generated PASCAL scores ($r = .62$), although PICSA was developed for longer public-speaking samples. The proportion of strongly prominent words correlated with both human ($r = .41$) and PASCAL-based charisma scores ($r = .47$), whereas articulation rate was associated only with the latter. Thus, systematic vocal-charisma judgments can arise from single short phrases, and local prominence provides relevant information even when temporal evidence for higher-order rhythmic organization remains sparse.

Keywords: *prominence, charisma, GRASS corpus, ASR, PICSA*

1. Introduction

Modern science and practice understands charisma as an attribution elicited by observable communicative signals. These signals convey qualities such as competence, confidence, and passion and comprise both verbal content and nonverbal delivery [1]. Acoustic-prosodic delivery is particularly influential: pitch, phrasing, vocal effort, as well as prominence-based word stress and rhythm systematically shape how charismatic a speaker sounds [2]–[5].

***Corresponding author**: olni@sdu.dk,*
**These two authors contributed equally to this paper and are in alphabetical order.*

Delivery also precedes content in perceptual processing. Listeners rapidly extract nonverbal information and use it to form initial charisma impressions, whereas lexical, syntactic, and rhetorical information require more time and may subsequently confirm or revise these impressions [6]. However, 'rapid" has rarely meant exposure to only a single short phrase. Previous experiments often used cumulative gating designs, or thin slices embedded in a wider range of stimulus durations. Longer spoken samples and phrases have been associated with higher charisma ratings [2], [7]. Ultra-short stimuli may therefore sound less charismatic; more fundamentally, it remains unclear whether excerpts of, e.g., 1–2 s, contain sufficient information to yield systematic judgments at all.

The same knowledge gap exists for automated assessment tools. The PICSA algorithm integrates 21 acoustic-prosodic features into a continuous PASCAL score of vocal charisma. Its criterion validity has been demonstrated across several domains [8]. PASCAL scores correlated with Danish students' (oral-exam) performances at individual and team levels ($r = .51$–$.59$) and with team performance in a creative task ($r = .70$); an earlier study accounted for 56% of the variance in individual oral-examination grades [9]. More directly, PICSA recently predicted human charisma ratings of synthetic voices with $r = .897$, explaining more than 80% of their variance [8]. Yet these applications used samples of approximately 30–40 s or longer, allowing the means and distributions of PICSA's acoustic parameters to stabilize. Whether comparable convergence is retained for spontaneous excerpts roughly one tenth as long is unknown.

Prosodic prominence is a candidate cue under these conditions. Unlike aggregate measures such as mean pitch level or speaking rate, it varies locally between syllables through changes in pitch, duration, intensity, and articulatory precision. Higher prominence levels thus constitute local vocal-effort pulses.

Previous charisma research has captured effort-related delivery at broader temporal scales in two ways. One line examined global, sample-level measures of intensity, voice quality, and articulatory precision independently of prominence, finding a positive linke between effort and charisma levels [3], [4]. Another examined sequences of prominence-related pulses through speech rhythm and amplitude modulation: rhythm measures correlated with perceived charisma ($\rho = .41$–$.43$) [5], while stronger, more regular amplitude modulations increased charisma ratings [10]. Thus, charisma has been linked to global effort-related settings and extended pulse sequences, but not to individual prominence pulses as cues in their own right. Since speech rhythm also includes slower phrasal organization [11], a single short inter-pausal unit provides only sparse evidence for such higher-order patterns.

The unresolved question is therefore whether local effort pulses already cue charisma before their sequence establishes broader effort patterns or stable rhythm. Informal spontaneous dialogue provides a particularly stringent test: compared with performed read or elicited speech, its prosodic shaping is less controlled and its generally modest vocal effort may reduce the salience of individual pulses.

Two scale-dependent predictions can be made. If individual prominence pulses cue charisma in their own right as local manifestations of vocal effort, excerpts with more prominent words—and particularly higher proportions of strongly prominent words—should sound more charismatic, whereas unaccented words should show the opposite relation. If such pulses become relevant only through their temporal organization effort levels or patterns such as rhythm, however, local counts and proportions should be weak predictors of charisma.

We test these predictions in 92 inter-pausal units of 6–9 words and 1–2 s duration from spontaneous Austrian German dialogues. Thirty listeners rated each excerpt for charisma, likeability, self-confidence, and passion. We ask whether (1) these ratings exhibit a coherent, stimulus-related structure at the group level; (2) human charisma ratings converge with PICSA-generated PASCAL scores based on broader acoustic-prosodic measures, despite PICSA's development on longer public-speaking samples; and (3) the number, strength, and relative proportion of local prominence pulses predict human and PASCAL-based charisma scores. Together, these questions probe the temporal lower boundary of human and algorithmic charisma assessment and test whether individual prominence pulses remain informative before extended rhythmic organization becomes available.

2. Charisma Perception Experiment

2.1. Stimuli

The empirical basis of the present study is the GRASS corpus, which provides high-quality recordings of both read and spontaneous German speech produced in a variety of communicative settings [12], [13]. For the purposes of the present study, we focus on spontaneous, conversational speech, as this domain captures natural interactional dynamics and thus represents a particularly relevant testing ground for the perception of charisma under ecologically valid conditions.

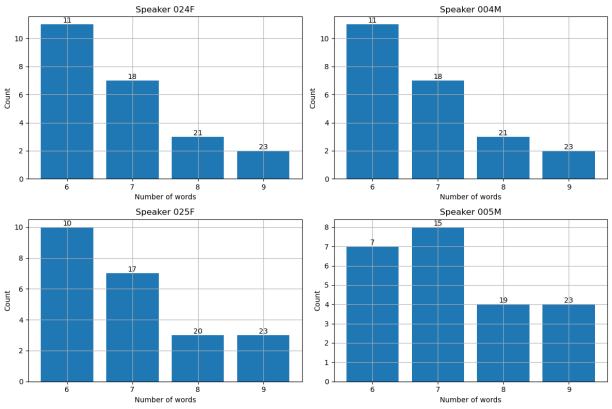


Figure 1: Distribution of the utterance lengths for the individual speakers. 23 utterances where selected for every speaker with number of words ranging from 6 to 9 per utterance.

From this corpus, a subset of short prosodic phrases, corresponding to inter-pausal units (IPUs), was extracted for two mixed-gender conversations (004M024F and 025F005M). These units were selected to represent a broad range of prominence patterns, including variation in both the number and the strength of prominences as annotated within the KIM framework. To isolate the effect of prominence on perceived speaker charisma, we applied the following filtering criteria: First, we included only stimuli containing at least six words. We then excluded utterances containing disfluencies, unintelligible segments, or laughter to remove confounding variables. Finally, we ensured a one-to-one correspondence between orthographic word tokens and prominence tokens. After the filtering, we obtained 23 stimuli for each speaker (i.e. 23x4 stimuli in total) with utterance lengths ranging from 6 to 9 words (see Figure 1 for the detailed distribution). The resulting dataset therefore allows for a controlled comparison of different prominence configurations while maintaining the natural variability characteristic of spontaneous speech.

2.2. Participants

Thirty native speakers of German residing in Austria were recruited via Prolific. Their mean age was 30.3 years ($SD = 9.5$, range: 18–57); 18 identified as male, 11 as female, and one did not report gender. Participants represented different educational and occupational backgrounds. All provided informed consent and were compensated in accordance with Prolific's standard procedures.

2.3. Procedure

The online perception experiment was conducted via SoSciSurvey. Participants completed the self-paced

task individually using their own listening equipment. They heard the 92 short excerpts from casual conversations, presented in individually randomized order, and could replay stimuli or take breaks as needed. After each excerpt, they rated the speaker's charisma, likeability, self-confidence, and passion on continuous scales from 1 to 100.

The selection of dimensions was guided by the 3-pillar model of vocal charisma comprising perceived competence, self-confidence, and passion [1]. Competence was omitted because its assessment relies partly on higher-order characteristics such as phrasing, pausing, and rhythmic organization. Likeability was included instead as a complementary dimension relevant to rapid social impressions of speakers [2], [14].

Participants were instructed to judge exclusively how the speaker sounded, irrespective of semantic content, and to respond spontaneously. They were informed that there were no correct or incorrect answers.

3. Analysis methods

3.1. Prosodic prominence annotations

All stimuli used in the experiment were annotated with word-level prominence level labels following the annotation guidelines presented in [12], using a three-level prominence scale (none=0, weak=1, strong=2). Since GRASS was manually prosodically annotated only on a subset of its speakers (i.e., also only a subset of the speakers of our stimuli set), we chose to follow a two-stage prominence annotation process for all stimuli of our experiment: First, we used a wav2vec2-based prominence detector to assign a prominence label to each word [15]. Subsequently, the fourth author of this publication manually corrected all automatically detected prominence levels to ensure annotation quality and consistency across the dataset.

3.2. Prosodic prominence features

Based on these labels, we calculated *prominence count features*, which capture the raw frequency of different prominence levels (none, weak, strong) within each utterance. Next, we computed *prominence ratio features* that normalize the *prominence count features* by the total number of prosodic words per utterance to provide relative distributions, including a contrast intensity measure that quantifies the balance between strong and weak prominence. The last category are the *prominence rate features* that normalize prominence counts by utterance duration to provide a measure that captures prominence density over time.

3.3. Forced-alignment and ASR-based features

Forced-alignment features provide temporal and pronunciation-based measures including articulation rate, word rate, and pronunciation deviation metrics (levenstein distance of the pronunciation of the stimulus compared to its canonical pronunciation). Additionally, we include the word error rate (WER) from a fine-tuned wav2vec2-based ASR system with lexicon and language model [16], as the WER may also serve as an indicator for the intelligibility of the utterance (combining pronunciation and content). Table 1 summarizes all prominence features (counts, ratios and rates) as well as features extracted from the forced-alignment/ASR features [16], [17].

4. Results

4.1. How do ratings of likeability, self-confidence and passion relate to perceived charisma in casual conversations?

Figure 2 presents pairwise correlation analyses between all the perceptual dimensions *likeability, self-confidence* and *passion* with *charisma*. Each data point represents one of the 92 speech stimuli, where coordinates correspond to mean participant ratings for that stimulus. All correlations are statistically significant at $p < 0.001$ (***). The analysis reveals two clusters of highly correlated dimensions: charisma shows strong positive correlations with likeability ($r = 0.897$) and moderate correlations with self-confidence ($r = 0.785$) and passion ($r = 0.794$). Additionally, self-confidence and passion are highly correlated ($r = 0.900$). Likeability exhibits the weakest correlations with self-confidence ($r = 0.592$) and passion ($r = 0.626$), suggesting it captures a somewhat distinct aspect of perceived charisma.

4.2. How do human charisma ratings correlate with PASCAL scores?

Figure 3 shows a correlation plot with the charisma ratings from the human participants (Rating01) as dependent variable, and the automatically derived PASCAL score for charisma (2nd column). We observe that both scores are approximately Gaussian distributed. However, the distribution of the PASCAL scores is noticeably broader, whereas the human charisma ratings exhibit a much sharper distribution. And yet, despite the substantially shorter duration of the stimuli used in the present experiment, PASCAL scores showed strong correlations with the listener ratings ($r = 0.62$, $p < 0.001$). This demonstrates a robust predictive validity of the PASCAL scores even under conditions of limited temporal input, and above that for spontaneous conversation excerpts. This result further lets us hypothesize that - as PASCAL is based on acoustic (prosodic) features of public speaking styles - the acoustic cues contributing to perceived charisma in casual speech need to be strongly overlapping with those cuing charisma in public speaking styles.

4.3. Is there a relationship between the prominence level distribution and perceived charisma?

In the following section, we focus in more detail on how the charisma relate to the acoustic properties of the casual conversational stimuli. For this we extracted a large set of features relating to prosodic prominence (how many words in an utterance are prominent to which degree), articulation rate, pronunciation distance from the canonical, and the WER. Among the prominence count features and prominence rate features shown in Table 1, the strongest correlation could

Table 1: Full set of prominence counts, ratios and rates, along with and forced-alignment and ASR-based features.

Feature name	Type	Description or mathematical formulation
Prominence count features		
feat_num_prom2	Discrete	Count of words with strong prominence (level 2)
feat_count_plvl0	Discrete	Count of prominence level 0 annotations
feat_count_plvl1	Discrete	Count of prominence level 1 annotations
feat_count_plvl2	Discrete	Count of prominence level 2 annotations
feat_count_plvl01	Discrete	Count of prominence levels 0 and 1 combined
feat_count_plvl12	Discrete	Count of prominence levels 1 and 2 combined
feat_count_plvl02	Discrete	Count of prominence levels 0 and 2 combined
Prominence ratio features		
feat_ratio_plvl0	Continuous	$\frac{\text{count\_plvl0}}{\text{total\_prominence\_levels}}$
feat_ratio_plvl1	Continuous	$\frac{\text{count\_plvl1}}{\text{total\_prominence\_levels}}$
feat_ratio_plvl2	Continuous	$\frac{\text{count\_plvl2}}{\text{total\_prominence\_levels}}$
feat_ratio_plvl01	Continuous	$\frac{\text{count\_plvl0+count\_plvl1}}{\text{total\_prominence\_levels}}$
feat_ratio_plvl12	Continuous	$\frac{\text{count\_plvl1+count\_plvl2}}{\text{total\_prominence\_levels}}$
feat_ratio_plvl02	Continuous	$\frac{\text{count\_plvl0+count\_plvl2}}{\text{total\_prominence\_levels}}$
feat_ratio_plvl2_to_plvl0	Continuous	$\frac{\text{count\_plvl2}}{\text{count\_plvl0+count\_plvl2}}$ (contrast intensity)
Prominence rate features		
feat_prom_rate	Continuous	$\frac{\sum \text{prominence\_levels}}{\text{duration}}$
feat_prom_rate_plvl0	Continuous	$\frac{\text{count\_plvl0}}{\text{duration}}$ (low prominence density)
feat_prom_rate_plvl1	Continuous	$\frac{\text{count\_plvl1}}{\text{duration}}$ (medium prominence density)
feat_prom_rate_plvl2	Continuous	$\frac{\text{count\_plvl2}}{\text{duration}}$ (high prominence density)
Forced-alignment/ASR features		
feat_dur	Continuous	Utterance duration in seconds
feat_AR	Continuous	$\frac{\text{num\_phones}}{\text{duration}}$ (articulation rate)
feat_wrd_rate	Continuous	$\frac{\text{num\_words}}{\text{duration}}$ (word rate)
feat_wer_w2vLM	Continuous	Word error rate from wav2vec2 ASR system
feat_prondiff	Continuous	Pronunciation difference score between used and longest possible pronunciation
feat_levdist	Continuous	Levenshtein distance between used and longest possible pronunciation

be found with the ratio of (level2) prominent words per utterance (i.e., feat_ratio_plvl2), both with the human charisma ratings ($r = 0.41$, $p < 0.001$), and with the PASCAL scores ($r = 0.47$, $p < 0.001$) (see Figure 3). A significant negative correlation but lower in effect was observed between the feature feat_num_prom0 and the perceived charisma score ($r = -0.24$, $p = 0.019$), showing that utterances with a higher numbers of completely unaccented words tend to result in lower degrees of perceived charisma.

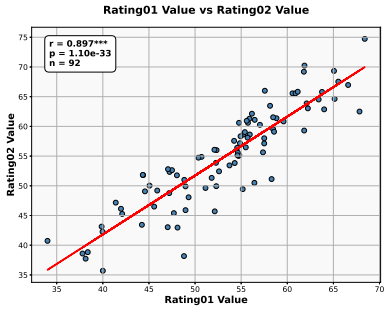
Contrary to our expectation, a pronunciation closer to the standard (less dialectal and less reduced), as reflected by the feature feat_prondiff, did not affect nor perceived nor automatically derived charisma scores. That PASCAL scores do not correlate to feat_prondiff might not come as a surprise, as its internal features do not include a knowledge base on Austrian German pronunciation variation. That human listeners are also not sensitive, however, comes as a surprise given sociolinguistic literature often relating standard pronunciation in Austria to be related to perceived competence.

Articulation rate, despite influencing PICSA scores, did not contribute to perceived charisma. Both human and PISCA scores were further also not significantly correlated to the utterance-level WER. This is actually a positive result when considering future work in which one might want full automatic processing pipelines (as sometimes they start with ASR, where then other derived metrics are biased by WER).

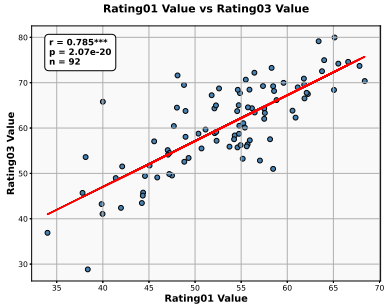
5. Discussion and Conclusions

The results show that 1–2 s excerpts of spontaneous conversation already support systematic group-level charisma judgments. Human ratings converged with PICSA-generated PASCAL scores, and the proportion of strongly prominent words, i.e. the number and strength of local vocal-effort pulses, predicted both human and algorithmic charisma assessments. The proportion of strongly prominent words correlated with human ($r = .41$) and PASCAL-based charisma scores ($r = .47$), whereas unaccented words showed a weaker negative association with human ratings.

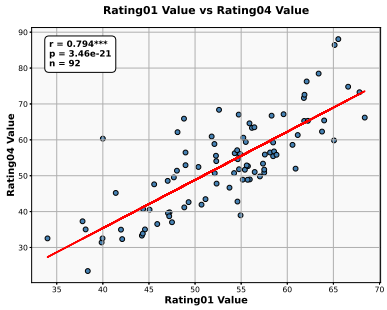
At the same time, the rating structure indicates that not all charisma components are equally accessible within such ultra-short intervals. Charisma was most strongly related to likeability ($r = .897$), whereas self-confidence and passion formed an equally close pairing ($r = .900$). This agrees with earlier links between charisma, charmingness, enthusiasm, and attractiveness [2], [14]. Longer performances, however, associate charisma particularly with visionary, inspiring, and self-confident impressions [18]. The difference may reflect temporal cue availability. Self-confidence



(a) Charisma vs likeability



(b) Charisma vs self-confidence



(c) Charisma vs passion

Figure 2: Pairwise correlation plots between the human *charisma ratings* with those of *likeability*, *self-confidence* and *passion*, given the mean rating values (N=92).

relies partly on intensity contrasts and phrase-final pitch movements, with one short IPU having at most one final contour. Passion likewise involves pitch and intensity variability as well as speaking rate, which require time to unfold. Competence ratings were omitted for the same reason in our study: pausing, phrasing, and temporal organization operate largely beyond a single prosodic phrase. Ultra-short stimuli may therefore foreground an initial socially evaluative form of charisma, while its dynamic and competence-related components become differentiated over longer intervals.

The human–PASCAL correlation ($r = .62$) further indicates that much of the acoustic information modeled by PICSA is locally available and generalizes from public speaking to casual conversation. However, PASCAL scores were more widely distributed than human ratings. Thus, PICSA captured the relative ordering of the excerpts, but phrase-level scores may require separate calibration because acoustic estimates are less stable in ultra-short time windows.

Within the observed range, the prominence relationships were approximately linear. However, acoustic charisma cues generally have feature-specific sweetspots beyond which further increases cease to be beneficial [1], [2]. Excessive accentuation should eventually weaken informational contrast and sound overemphatic. The present phrases apparently did not reach this upper boundary, making the optimal prominence density and its possible reversal important targets for future research.

WER was unrelated to either charisma measure. Although ASR errors can partly reflect pronunciation reduction and articulatory effort, all excerpts represented informal spontaneous speech with a broadly moderate degree of reduction. This restricted range, together with the instability of WER for 6–9-word samples, may have obscured such effects. The null result therefore does not imply that articulatory effort is irrelevant, but that utterance-level WER did not differentiate it reliably here.

Finally, the stimuli came from only four speakers, and naturally occurring prominence may covary with speaker identity and information structure. Larger speaker samples and controlled manipulations of prominence density, spacing, and stimulus duration are therefore required. Overall, vocal charisma in a few seconds already has a coherent perceptual and acoustic basis, but its dimensional composition depends on how much time the voice is given to unfold.

6. Acknowledgments

This work resulted from a Visiting Guest Professorship from O. Niebuhr at Graz University of Technology, granted by the Faculty of Electrical and Information Engineering. Participants were kindly paid for their rating task by AllGoodSpeakers ApS; also note that the 2nd author the CEO of that company, please see https://oliver niebuhr.com/conflict-of-interest.html for a COI statement.

References

[1] J. Michalsky and O. Niebuhr, "Myth busted? challenging what we think we know about charismatic speech," *AUC Philologica*, vol. 2019, no. 2, pp. 27–56, 2019.

[2] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, vol. 51, no. 7, pp. 640–655, 2009.

[3] O. Niebuhr and R. Skarnitzl, "Those who shout the loudest—do they sound more charismatic?" In *Proc. 1st Int. Conf. on Tone and Intonation (TAI)*, 2021, pp. 1–5.

[4] S. Berger, ""Like, comment, subscribe"—perception of acoustic-prosodic features of content creators' charismatic speech on YouTube," Ph.D. dissertation, Kiel University, Kiel, Germany, 2024.

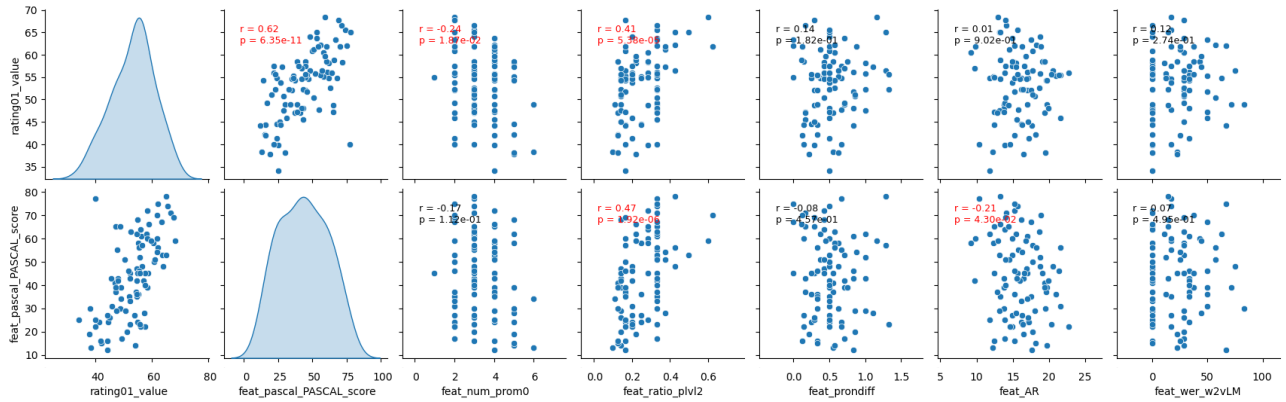


Figure 3: Correlation plots showing how the charisma ratings from the human participants (Rating01) relate to the automatically derived PASCAL score for charisma, and to the derived prominence- , articulation rate-, and ASR-related features.

[5] O. Niebuhr and N. Taghva, "How rhythm metrics are linked to produced and perceived speaker charisma," in *Proc. 25th Interspeech Conference*, 2024, pp. 1065–1069.

[6] A. Caspi, R. Bogler, and O. Tzuman, "Judging a book by its cover: The dominance of delivery over content when perceiving charisma," *Group & Organization Management*, vol. 44, no. 6, pp. 1067–1098, 2019.

[7] O. Jokisch, V. Iaroshenko, M. Maruschke, and H. Ding, "Influence of age, gender and sample duration on the charisma assessment of German speakers," in *Proc. 29th ESSV*, Ulm, Germany, 2018, pp. 224–231.

[8] L. Conle, Ï. Valls-Ratés, and O. Niebuhr, "Synthetic yet striking? assessing vocal charisma in TTS via perceptual and algorithmic measures," in *Proc. IEEE ICASSP*, 2026, pp. 17 537–17 541.

[9] O. Niebuhr, "Advancing higher-education practice by analyzing and training students' vocal charisma: Evidence from a Danish field study," in *Proc. 7th Int. Conf. on Higher Education Advances (HEAd'21)*, 2021, pp. 743–751.

[10] H. R. Bosker, "The contribution of amplitude modulations in speech to perceived charisma," in *Voice Attractiveness: Prosody, Phonology and Phonetics*, B. Weiss, J. Trouvain, M. Barkat-Defradas, and J. J. Ohala, Eds., Singapore: Springer, 2021, pp. 165–181.

[11] A. V. Barchet, M. J. Henry, C. Pelofi, and J. M. Rimmele, "Auditory-motor synchronization and perception suggest partially distinct time scales in speech and music," *Communications Psychology*, vol. 2, no. 1, p. 2, 2024.

[12] B. Schuppler, M. Hagmüller, and A. Zahrer, "A corpus of read and conversational Austrian German," *Sp. Comm.*, vol. 94, pp. 62–74, 2017.

[13] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: The Graz corpus of Read And Spontaneous Speech," in *Proc. LREC*, 2014, pp. 1465–1470.

[14] R. Signorello, F. D'Errico, I. Poggi, and D. Demolin, "How charisma is perceived from speech: A multidimensional approach," in *Proc. of IEEE SocialCom-PASSAT*, 2012, pp. 435–440.

[15] J. Linke and B. Schuppler, *Prominence-aware automatic speech recognition for conversational speech*, 2025. arXiv: `2509.10116 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2509.10116`.

[16] "What's so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures," *Computer Speech & Language*, vol. 90, p. 101 738, 2025, ISSN: 0885-2308.

[17] J. Linke, S. Wepner, G. Kubin, and B. Schuppler, *Using Kaldi for Automatic Speech Recognition of Conversational Austrian German*, 2023. arXiv: `2301.06475 [cs.CL]`. [Online]. Available: `https://arxiv.org/abs/2301.06475`.

[18] F. Weninger, J. Krajewski, A. Batliner, and B. Schuller, "The voice of leadership: Models and performances of automatic analysis in online speeches," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 496–508, 2012.