Production Strategies of Vocal Attitudes

Léane Salais, Pablo Arias, Clément Le Moine, Victor Rosi, Yann Teytaut, Nicolas Obin, Axel Roebel

Interspeech 2022
18-22 September 2022, Incheon, Korea

Production Strategies of Vocal Attitudes

Léane Salais[*1], *Pablo Arias*[*1,2], *Clément Le Moine*[*1], *Victor Rosi*[1], *Yann Teytaut*[1], *Nicolas Obin*[1], *Axel Roebel*[1]

[1] STMS Lab - IRCAM, Sorbonne Université, CNRS, Ministère de la Culture, Paris, France
[2] Lund University Cognitive Science, Lund University, Lund, Sweden

Abstract

Humans have an impressive ability to communicate precise social intentions and desires with their voice — through vocal attitudes. Previous studies have shown how isolated acoustic features such as pitch can convey social attitudes, but have mostly worked with single attitudes and have not controlled for inter-speaker variability. Thus, the vocal behaviours used to produce social attitudes remain mostly unknown. That is the aim of the current study, to uncover the anatomic production strategies that speakers use to communicate vocal attitudes. To do this, we analysed recordings from N=20 French speakers producing dominant, friendly, seductive and distant speech. For each of these attitudes, we investigated their vocal fold behaviour, vocal tract actuation and phonetic speech structure, with the support of deep alignment methods, and compared them with group statistics. We notably produced high-level representations of speakers' articulation (e.g. Vowel Space Density) and speech rhythm. Our results reveal speakers' prototypical strategies to produce vocal attitudes, and highlight how vocal behaviours can communicate social signals. We expect these results to provide an objective validation method for deep voice attitude conversions.

Index Terms: vocal social attitudes, speech production, articulation

1. Introduction

Human interactions are governed by a fantastic interplay of social signals. At the heart of this mechanism lies our ability to communicate social attitudes with interacting partners [1]. For example, a person can be friendly, distant, seductive or dominant with a stranger they just met, depending on the outcome they expect from the interaction. Such attitudes differ from emotions, because they do not only hint at the speakers' affective state, but at their social intention [2]. However, despite the ubiquity of vocal attitudes in social contexts, it remains mostly unknown how they are communicated vocally. In this work, we present an anatomically based acoustic evaluation of speech utterances to understand how speakers modulate their speech to communicate social attitudes.

Previous studies have provided insights on how to decode attitudes from vocal signals. Cognitive psychology studies have identified acoustic signatures used by listeners to infer attitudes from speech [3], e.g. showing how pitch can affect the general perception of dominance [4] or friendliness [5]. In parallel, speech processing studies have developed automatic vocal attitude detection methods, and identified the most important features in multi-attitude datasets [6, 7, 8]. However, to our knowledge, previous studies did not investigate vocal attitudes from their underlying anatomical mechanisms, thus preventing a holistic understanding of both the speakers' ongoing cognitive processes and their social intentions. Moreover, they usually do not account for multi-level temporal variations in speech, e.g. at the phoneme scale. Thus, the specific vocal strategies used by speakers to produce social attitudes remain mostly unknown.

Interestingly, acoustic analysis techniques can provide quantitative descriptions of the anatomical mechanisms happening during speech production. First, we can estimate the amount of energy and expressiveness a speaker puts into their vocalisations by studying their vocal fold behaviour — e.g. their vocal loudness, vocal fold saturation and vocal pitch contours. Second, we can assess a speaker's articulation strategies by studying their vocal tract actuation — e.g. measuring formant frequencies and deducing Vowel Space Densities [9]. Finally, with a phoneme-to-audio automatic alignment [10], we can describe phoneme-scale prosodic modulations — e.g. by focusing on rhythmic speech patterns.

In this work, we aim to give a clear account of these processes in the context of vocal attitude production. To do this, we analysed a multi-speaker French speech database [11] showcasing four social attitudes (friendliness, dominance, seductiveness and distance), and investigated the impact of these attitudes on the way speakers control their (1) vocal fold behaviour, (2) vocal tract actuation and (3) phonetic speech structure. In the following, we detail the acoustic descriptors used to analyse voice production mechanisms, and deploy group statistic methods to uncover speakers' attitude production strategies (Fig. 1).

We hence present two major contributions: 1) an anatomically based acoustic evaluation method of vocal attitudes, derived from state-of-the-art phoneme-to-audio alignments, and 2) its direct application to uncover the prototypical vocal strategies underlying social attitudes communication.

2. Data extraction

2.1. Att-HACK : a dataset for speech attitudes

The Att-HACK dataset [11] features 30 hours of expressive speech in French, with 20 speakers (9 men and 11 women) portraying four social attitudes (friendly, seductive, dominant and distant) over 100 isolated sentences. Att-HACK is freely available for research under a Creative Commons (BY/NC/ND) License. All the audio signals are provided with their orthographic text transcription. Each utterance is recorded in three to five versions for a given speaker and attitude, allowing us to glance at multiple production strategies from the same person.

In this study, we used a subsample of the Att-HACK database. We randomly sampled two recordings per speaker and per attitude for 62 sentences, thus obtaining 2400 recordings per attitude. The 62 sentences were selected to maximise semantic diversity, i.e. achieve an optimal coverage of the semantic space yielded by the CamemBERT [12] French language model.

2.2. Phoneme-to-audio alignments

To investigate articulatory and phonetic vocal strategies, we need to access the segmental information in speech (i.e. tem-

10.21437/Interspeech.2022-10947

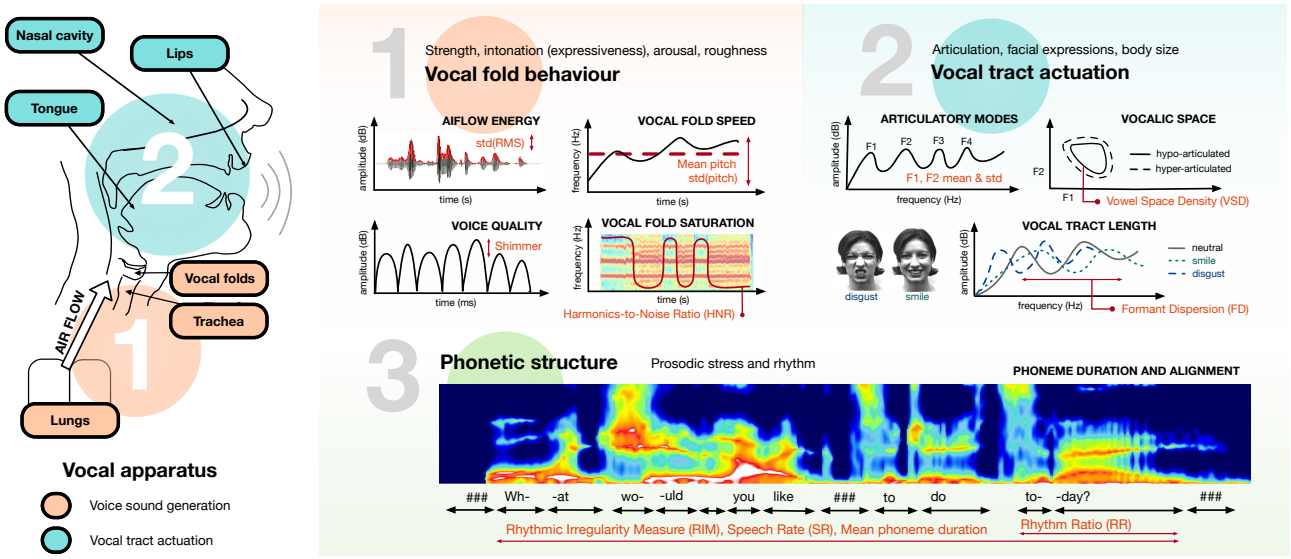

Figure 1: *Anatomical voice production mechanisms and corresponding acoustic features. To describe the production strategies of vocal attitudes, we analyse three main categories of speech descriptors, relating to (1) vocal fold behaviour, (2) vocal tract actuation and (3) phonetic structure.*

poral information at the phoneme level). To infer it from Att-HACK recordings, we generate phoneme-to-audio alignments using a recent deep learning-based phonetic aligner [10] .

To do this, we generated phonetic sequences and mel-spectrograms from recordings. We first process the text with a phonemizer [13] to convert words into the French Phonetic Alphabet (FPA), and add pause tokens around the obtained phone sequences to account for silence sections. We then process the audio into normalised, log-scaled mel-spectrograms. Given these inputs, the model learns to predict relevant posteriograms, i.e. the per-frame probability distribution over all phonemes, by minimising a constrained Connectionist Temporal Classification (CTC) loss [14]. Since we do not seek generalisation to any kind of unseen data, we train the model to overfit on the Att-HACK dataset, and derive forced alignment from the posteriograms following the procedure in [10].

To assess the alignment obtained on Att-HACK, we compare it with the ground truth for the multi-speaker TIMIT dataset [15]. We know that the chosen aligner is able to align TIMIT with a state-of-the-art average precision of 16.3ms [10]. Given that the longest phoneme in TIMIT lasts 605ms, phonemes with an even longer predicted duration in Att-HACK must be misaligned. There are 1698 such phonemes, leading to an estimate of 0,4% of drastic errors.

3. Anatomic voice production assessment

We built our analysis on a list of selected acoustic variables, following previous research on emotional speech [16]. We split them into three clusters that reflect the aforementioned voice production mechanisms (Fig. 1).

3.1. Vocal fold behaviour (Fig. 1-1)

To quantify speakers' control over their vocal folds, we use acoustic descriptors of their vibration amplitude and rate.

First, we estimate the voice signal's Root Mean Square (RMS, dB) and its standard deviation (window size=2048). Although RMS reflects the general airflow energy—which also affects whispers and unvoiced phonemes — here we focus on its impact on vocal folds vibration strength.

Second, we measure the Harmonics-to-Noise Ratio (HNR) as an indicator of vocal fold saturation in vocalisations. In speech, HNR measures the energy ratio between harmonics produced by the vibrating vocal folds and the glottal noise in the spectrum. A sustained and subtle airflow produces harmonic vocal fold vibrations with a high HNR; in contrast, strong airflow from the lungs makes the vocal folds oscillate in non-linear or chaotic regimes, resulting in a rough voice [17]. This feature has been previously linked with e.g., aversiveness, arousal, negative valence and to some extent, emotion intensity [18].

As a complementary measure of HNR we measure Shimmer (dB), which is associated with voice quality. Shimmer is the voice's amplitude variation over glottis cycles: high shimmer is often associated with a breathy voice.

Finally, we measure vocal pitch (Hz, mean and standard deviation), which reflects the vocal fold's vibration speed, i.e. the count of glottis cycles. Its variations summarise the modulations in intonation, a key feature to communicate vocal intentions, attitudes and emotions [19, 20, 21].

3.2. Vocal tract actuation (Fig. 1-2)

To investigate vocal articulation strategies, we also extracted vocal spectral features. Specifically, we measured the first and second formant frequencies (F1, F2, Hz, mean and standard deviation [22]), which represent the articulatory resonances of the vocal tract: they are impacted by the lips, mouth and tongue positions. Formants are not only essential to convey phonetic information, but also key to convey emotional information such as facial expressions [23].

To estimate the speakers' dynamic vocal tract elongation, we measure Formant Dispersion (FD) (Hz) i.e. the averaged difference between successive formant frequencies (F1 to F4). FD reflects the vocal tract length — which is also closely tied to body size [24]. Speakers can extend (lower FD) or shorten (higher FD) their vocal tract [25] through facial expressions: previous results have reported an association between FD and

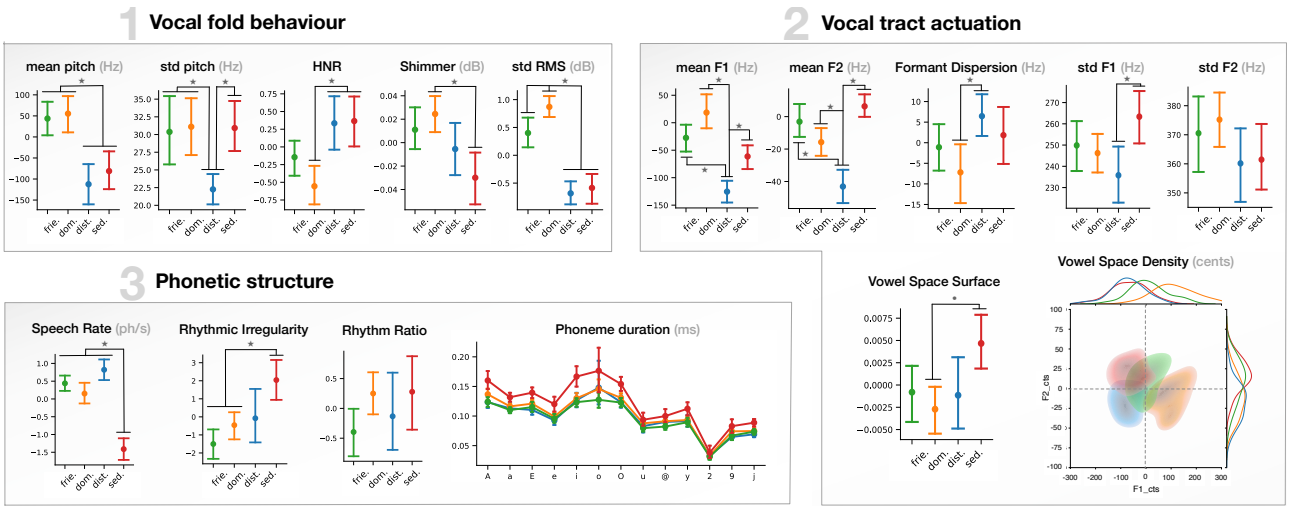

Figure 2: *Feature analyses for (1) vocal fold behaviour, (2) vocal fold actuation and (3) phonetic structure on friendliness (green), dominance (orange), distance (blue) and seductiveness (red). '⋆' : statistically significant difference ($p<0.05$), '•' : marginally significant difference ($p<0.1$); paired t-tests. Error bars represent 95% confidence intervals on the mean.*

expressions of emotions such as smiles [26] and disgust [27]. However, FD does not allow for an exhaustive description of the underlying articulatory strategies, e.g. switching from one articulatory mode to another by shifting only one formant [28].

To accurately account for those strategies, we examine the vocalic space (VS), i.e. the space formed by F1 and F2 formants). We consider each vowel-related time frame in the dataset — extracted using the alignments described in Section 2.2. To study the topology of this space, we compute the Vowel Space Density (VSD) as proposed in [9]. We first estimate a probability density function for the count of time frames located in the neighbourhood of each point in the space, and normalise the density to [0, 1] for each speaker and attitude. To account for prototypical strategies, we only keep samples located in high density areas (above a threshold of 0.5). VSD offers a holistic understanding of vocal articulatory strategies. Positions of attitude clusters in the vocalic space reflect how speakers articulate to convey attitudes (articulatory modes, e.g., closed/open mouth), while the surface covering all samples in the VS represents how much they articulate: the broader the surface, the easier it is to discriminate the vowels pronounced [9].

3.3. Phonetic structure (Fig. 1-3)

To investigate speech's phonetic structure, we take advantage of the phoneme-to-audio alignments presented in Section 2.2. We estimate the Speech Rate (SR) — i.e. the mean number of phonemes per second in a speech utterance—, the Rhythmic Irregularity Measure (RIM) and the Rhythm Ratio (RR) [29]. The Rhythmic Irregularity Measure quantifies the mean duration difference between all segments in a sentence, whereas Rhythm Ratio is the mean duration difference between contiguous speech segments. These features yield indices on the global and local stability of speech rate.

4. Results

4.1. Statistical analysis

To statistically evaluate the differences in vocal production strategies, we analysed acoustic features with GLMMs (Generalised Linear Mixed Models). We report p-values, estimated from hierarchical model comparisons using likelihood ratio tests [30], and only present models that satisfy the assumption of normality (validated by visually inspecting the plots of residuals against fitted values) and statistical validation (significant difference with the nested null model). To test for main effects, we compare models with and without the fixed effect of interest. We perform post-hoc comparisons with paired t-tests, and apply Bonferroni corrections to correct for multiple comparisons. We report Cohen-d as a measure of effect size.

For each attitude, we present mean values of acoustic descriptors over full utterances. Because we are not investigating inter-speaker variability, but the speakers' own production strategies, we normalise features by speaker and get zero-centred values. Thus, variations between the conditions below reflect intra-speaker variations. In consequence, the statistical differences between attitudes bring out the shared part of the attitude production strategies among the speakers.

4.2. Vocal fold behaviour (Fig. 2-1)

We found a main effect of attitude on mean pitch ($\chi^2(3)=560$, $p<.001$), std pitch ($\chi^2(3)=396$, $p<.001$), HNR ($\chi^2(3)=83$, $p<.001$), shimmer ($\chi^2(3)=59$, $p<.001$) and std RMS ($\chi^2(3)=905$, $p<.001$). Post-hoc analyses revealed that mean pitch was higher for dominance and friendliness as compared to distance and seductiveness (paired t-tests, $p<.001$, $d>0.75$). Speakers' pitch variability was also smaller for distance than for other attitudes ($p=.05$, $d>0.9$). On another line, friendliness and dominance seemed to be opposed to seductiveness and distance in terms of dynamics and roughness. HNR was significantly higher in dominant speech than in distant ($p=.002$, $d=0.68$) and seductive speech ($p=.001$, $d=0.7$); similarly, we found higher RMS variability for friendliness and dominance as compared to distance and seductiveness ($p<.001$, $d>1.1$). In addition, shimmer was significantly higher for dominant utterances as compared to seductive ones ($p=.001$, $d=0.7$).

4.3. Vocal tract actuation (Fig. 2-2)

We found a main effect of attitude for Formant Dispersion (FD) ($\chi^2(3)=61$, $p<.001$), F1 ($\chi^2(3)=99$, $p<.001$), F2 ($\chi^2(3)=37$,

p<.001), std F1 (χ^2(3)=73, p<.001) and std F2 (χ^2(3)=24, p<.001). Post-hoc analyses revealed that speakers significantly decreased their FD when producing dominance as compared to distance (p=.02, d=0.6). In line with this result, we found significantly lower F1 (p=.005, d>0.8) and F2 (p<.001, d>0.8) frequencies for distance, compared to all other attitudes. On another line, we found that distant utterances were produced with significantly more F1 variability as compared to seductiveness (p=.01, d=1.2), but found no significant differences for std F2. Finally, we only found a marginal difference between the surfaces of VSDs of the attitudes between seductive and dominant attitudes (p=0.06; see VSD plot in Fig. 2).

4.4. Phonetic structure (Fig. 2-3)

We found a main effect of attitude for Speech Rate (SR) (χ^2(3)=81, p<.001) and Rhythmic Irregularity (RIM) (χ^2(3)=18, p<.001), but no significant effect on Rhythm Ratio (RR) (χ^2(3)=4.2, p=.23). That is, attitudes influence global rhythmic patterns rather than local ones.

Post-hoc analyses revealed that seductive samples had a significantly lower SR (p<.001, d>1.24) and higher RIM as compared to friendliness and dominance (p=.02, d>0.7). The duration of all vowels was also extended accordingly.

5. Discussion

In the present study, we investigated how speakers modulate their voice to communicate vocal attitudes. To do this, we analysed the vocal production of dominant, friendly, seductive and distant attitudes in a multi-speaker and multi-attitude French database. For each attitude, we reported the changes in the speakers' vocal fold behaviour, vocal tract actuation, and phonetic speech structure.

In line with previous findings, we found that dominance was expressed through a vocal tract elongation (lower formant dispersion) [31, 4] as well as a rough and dynamic voice (low HNR, shimmer, high std RMS). However, contrary to previous findings, speakers raised their pitch in comparison with other attitudes [32, 4]. This discordance may be explained by the language setting, culturally learned vocal associations, or more simply by the fact that previous studies contrasted dominance with neutral speech and not other vocal attitudes [19].

We also found strong prototypical strategies for seductiveness, which was produced with low pitch, low dynamics (low std RMS), and a relatively high harmonic content (high HNR, low shimmer) in comparison to other attitudes. Importantly, we also found a strong effect of seduction on speech's phonetic structure. Specifically, seductive utterances were produced with a slow and irregular rhythm, as if speakers took time to expose their intentions. Previous findings on vocal attractiveness report a lowered pitch for male speakers [33, 31, 34]. We complement these findings by studying seduction as a modulated vocal attitude, rather than an intrinsic vocal trait, and highlight the specific modulations speakers use to convey seductiveness.

Friendly prototypes were less marked than other attitudes in statistical terms. In any case, friendliness was produced with a raised and dynamic voice (high pitch, high std RMS). The speed and regularity of friendly versus seductive speech (higher SR, lower RIM) may hint at an uncomplicated and extraverted persona [35]. These results are in line with cross-lingual literature for English, Dutch, Chinese and Swedish [5, 36, 37].

The production strategies of distance were of particular interest. Indeed, distance was conveyed by fast speech that lacks expressiveness (low and steady pitch, high SR vs. seductiveness), pronounced with a low mouth aperture and a shortened vocal tract (low F1 and F2; high FD compared to dominance). In light of these results, it seems that when producing distance, speakers do not put much effort into being understood. Their calmness (e.g. high HNR when compared to dominance) suggests that their rendition of distance is close to indifference. Distance is hence distinct from neutrality, and could be interpreted as a marker of dissent, mistrust, or disgust.

More generally, to our knowledge, this is the first study to reveal diverging voice production strategies at the articulatory level. Specifically, we found that speakers' productions were distributed across specific clusters in the vowel space (Fig. 2-2). For example, we found that distance had a lower F1 than other attitudes, suggesting that distance is produced with a more closed mouth. Similarly, analysing the Vowel Space Density surface revealed that some attitudes span more articulatory modes than others. For instance, the vowel space for seductiveness was marginally wider than for dominance, which, in complement with formant dispersion findings, suggests that speakers switched between articulatory modes to produce vocal attitudes, by e.g, restraining or modulating their articulatory range. This result suggests that subtle cues in speech articulation can be used as a communicative signal of vocal intent.

Overall, these results shed light on the social intentions behind the production of social attitudes. For example, speakers limited their vocal expressivity to sound distant and hinted at a larger body size to sound dominant. Such behaviours may be closely interpreted from a social perspective, revealing the links between attitude-specific vocal behaviours and higher-order cognitive mechanisms [3]. However, it is important to highlight that the vocalisations analysed herein were produced by actors, and actors' vocalisations are known to be less authentic than spontaneous ones [38] — which, in the case of e.g. facial expressions of emotions, even seem to rely on different neural bases [39]. In any case, these results uncover the shared strategies used by speakers to volitionally produce vocal attitudes.

6. Conclusion

We showed that French speakers share common production strategies to communicate vocal attitudes such as friendliness, dominance, seductiveness or distance. To do this, we used speech descriptors and group statistics to uncover quantitative prototypes reflecting the speakers' vocal apparatus control.

Subsequent works will touch upon the interactions between the speaker's identity (e.g. gender) and the listeners' perception of these vocal attitudes.

On another line, this study is — to our knowledge — the first to use a deep phoneme-to-audio alignment model to investigate the articulatory quality and phonetic structure of speech. This method is fully replicable on any speech data. In future work, we plan to assess the reliability of deep semi-supervised voice attitude conversion models by comparing their outputs with the obtained prototypes.

7. Acknowledgements

This work was supported by the Sorbonne Université - Emergence project REVOLT (N.O., P.A.) and the French Ph2D/IDF MoVE project (C.L.M.).

8. References

[1] P. McAleer, A. Todorov, and P. Belin, "How do you say 'hello'? personality impressions from brief novel voices," *PloS one*, vol. 9, no. 3, p. e90779, 2014.

[2] A. Wichmann, "The attitudinal effects of prosody, and how they relate to emotion," *Proceedings of the ISCA Workshop on Speech and Emotion*, 01 2000.

[3] L. Goupil, E. Ponsot, D. Richardson, G. Reyes, and J.-J. Aucouturier, "Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature," *Nat. Commun.*, vol. 12, no. 1, p. 861, 2021.

[4] D. Puts, C. Hodges-Simeon, R. Cardenas, and S. Gaulin, "Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men," *Evolution and Human Behavior*, vol. 28, pp. 340–344, 09 2007.

[5] F. Chen, A. Li, H. Wang, T. Wang, and Q. Fang, "Acoustic analysis of friendly speech," vol. 1, 06 2004, pp. I – 569.

[6] F. Rosis, A. Batliner, N. Novielli, and S. Steidl, ""you are sooo cool, valentina!" recognizing social attitude in speech-based dialogues with an eca," vol. 4738, 09 2007, pp. 179–190.

[7] R. Ranganath, D. Jurafsky, and D. A. McFarland, "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates," *Computer Speech & Language*, vol. 27, no. 1, pp. 89–115, 2013, special issue on Paralinguistics in Naturalistic Speech and Language.

[8] A. Barbulescu, R. Ronfard, and G. Bailly, "Which prosodic features contribute to the recognition of dramatic attitudes?" *Speech Communication*, vol. 95, pp. 78–86, Dec. 2017.

[9] B. H. Story and K. Bunton, "Vowel space density as an indicator of speech performance," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. EL458–EL464, 2017.

[10] Y. Teytaut and A. Roebel, "Phoneme-to-audio alignment with recurrent neural networks for speaking and singing voice," *Proceedings of Interspeech 2021*, pp. 61–65, 2021.

[11] C. Le Moine and N. Obin, "Att-HACK: An Expressive Speech Database with Social Attitudes," in *Speech Prosody*, Tokyo, Japan, May 2020.

[12] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a tasty French language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7203–7219.

[13] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[15] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[16] P. Arias, L. Rachman, M. Liuni, and J.-J. Aucouturier, "Beyond correlation: Acoustic transformation methods for the experimental study of emotional voice and speech," *Emotion Review*, vol. 13, no. 1, pp. 12–24, 2021.

[17] K. A. Erhard, S. A. Kotz, E. Pfeifer, M. Besson, A. D. Friederici, and J. Matiasek, "On the relations of semantic and acoustic properties of emotions," in *In: Proceedings of the XIVth International Congress of Phonetic Sciences*, 1999, pp. 2121–2124.

[18] A. Anikin, K. Pisanski, and D. Reby, "Do nonlinear vocal phenomena signal negative valence or high emotion intensity?" *Royal Society open science*, vol. 7, no. 12, p. 201306, 2020.

[19] E. Ponsot, J. J. Burred, P. Belin, and J.-J. Aucouturier, "Cracking the social code of speech prosody using reverse correlation," *Proceedings of the National Academy of Sciences*, vol. 115, no. 15, pp. 3972–3977, 2018.

[20] L. Rachman, M. Liuni, P. Arias, A. Lind, P. Johansson, L. Hall, D. Richardson, K. Watanabe, S. Dubal, and J.-J. Aucouturier, "David: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech," *Behavior Research Methods*, vol. 50, no. 1, pp. 323–343, Feb. 2018.

[21] E. A. Piazza, M. C. Iordan, and C. Lew-Williams, "Mothers consistently alter their unique vocal fingerprints when communicating with infants," *Current Biology*, vol. 27, no. 20, pp. 3162–3167, 2017.

[22] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2011.

[23] P. Arias, C. Soladie, O. Bouafif, A. Roebel, R. Seguier, and J.-J. Aucouturier, "Realistic transformation of facial and vocal smiles in real-time audiovisual streams," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 507–518, 2018.

[24] A. Anikin, K. Pisanski, and D. Reby, "Static and dynamic formant scaling conveys body size and aggression," *Royal Society Open Science*, vol. 9, no. 1, p. 211496, 2022.

[25] M. Belyk, S. Waters, E. Kanber, M. Miquel, and C. Mcgettigan, "Individual differences in vocal size exaggeration," *Scientific Reports*, vol. 12, 02 2022.

[26] A. Drahota, A. Costall, and V. Reddy, "The vocal communication of different kinds of smile," *Speech Communication*, vol. 50, pp. 278–287, 04 2008.

[27] C. S. Chong, J. Kim, and C. Davis, "Disgust expressive speech," *Speech Commun.*, vol. 98, no. C, p. 68–72, apr 2018.

[28] K. Pisanski, A. Anikin, and D. Reby, "Vocal size exaggeration may have contributed to the origins of vocalic complexity," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 377, 01 2022.

[29] D. Gibbon and U. Gut, "Measuring speech rhythm," 01 2001, pp. 95–98.

[30] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[31] D. Feinberg, B. Jones, A. Little, D. Burt, and D. Perrett, "Manipulation of fundamental and formant frequencies influence the attractiveness of human male voices," *Animal Behaviour*, vol. 69, pp. 561–568, 03 2005.

[32] D. A. Puts, S. J. Gaulin, and K. Verdolini, "Dominance and the evolution of sexual dimorphism in human voice pitch," *Evolution and Human Behavior*, vol. 27, no. 4, pp. 283–296, 2006.

[33] S. Collins, "Men's voices and women's choices," *Animal behaviour*, vol. 60, pp. 773–780, 01 2001.

[34] Y. Xu, A. Lee, W. L. Wu, X. Liu, and P. Birkholz, "Human vocal attractiveness as signaled by body size projection," *PloS one*, vol. 8, p. e62397, 04 2013.

[35] F. Mairesse, M. Walker, M. Mehl, and R. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *J. Artif. Intell. Res. (JAIR)*, vol. 30, pp. 457–500, 09 2007.

[36] A. jun Li and H. Wang, "Friendly speech analysis and perception in standard chinese," in *INTERSPEECH*, 2004.

[37] D. House, "Phrase-final rises as a prosodic feature in wh-questions in swedish human-machine dialogue," *Speech Communication*, vol. 46, pp. 268–283, 07 2005.

[38] A. Anikin and C. F. Lima, "Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations," *Q. J. Exp. Psychol. (Hove)*, p. 17470218.2016.1, Jan. 2017.

[39] D. Valente, A. Theurel, and E. Gentaz, "The role of visual experience in the production of emotional facial expressions by blind people: a review," *Psychonomic bulletin & review*, vol. 25, 06 2017.