Production Strategies of Vocal Attitudes

Léane Salais*1, Pablo Arias*1,2, Clément Le Moine*1, Victor Rosi1, Yann Teytaut1, Nicolas Obin3, Axel Roebel1

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 method of vocal attitudes, derived from state-of-the-art phoneme-to-audio alignments, and 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-
Vocal apparatus

Voice sound generation

Vocal tract actuation

Nasal cavity
Lips
Tongue
Vocal folds
Teeth
Lungs

1. Strength, intonation (expressiveness), arousal, roughness

Vocal fold behaviour

AFLOW ENERGY
F1
F2

Hypo-articulated
Hyper-articulated

2. Articulation, facial expressions, body size

Vocal tract actuation

ARTICULATORY MODES

VOCALIC SPACE

3. Phonetic structure

Prosodic stress and rhythm

PHONEME DURATION AND ALIGNMENT

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.

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...
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 forms). 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 phone-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-meaning 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 ($p=0.9$, $d>0.9$). 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 ($p=0.02$, $d=0.68$) and seductive speech ($p=0.01$, $d=0.7$); 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 ($p=0.01$, $d=0.7$).

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 ($\chi^2(3)=37$, $p<.001$), and std F1 ($\chi^2(3)=50$, $p<.001$), std F2 ($\chi^2(3)=49$, $p<.001$) but not for $\chi^2(3)=47$, $p<.001$)}.
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=.05, 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=.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


