Att-HACK: An Expressive Speech Database with Social Attitudes

Clément Le Moine, Nicolas Obin

Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Att-HACK: An Expressive Speech Database with Social Attitudes

Clément Le Moine, Nicolas Obin

STMS Lab - IRCAM, CNRS, Sorbonne Université
Paris, France

Abstract

This paper presents Att-HACK, the first large database of acted speech with social attitudes. Available databases of expressive speech are rare and very often restricted to the primary emotions: anger, joy, sadness, fear. This greatly limits the scope of the research on expressive speech. Besides, a fundamental aspect of speech prosody is always ignored and missing from such databases: its variety, i.e. the possibility to repeat an utterance while varying its prosody. This paper represents a first attempt to widen the scope of expressivity in speech, by providing a database of acted speech with social attitudes: friendly, seductive, dominant, and distant. The proposed database comprises 25 speakers interpreting 100 utterances in 4 social attitudes, with 3-5 repetitions each per attitude for a total of around 30 hours of speech. The Att-HACK is freely available for academic research under a Creative Commons Licence.

Index Terms: expressivity, speech prosody, social attitude, speech database

1. Introduction

Though the linguistic functions of speech prosody are nowadays well documented in a large number of languages (phonology, syntax/prosody interface, etc...), its expressive or para-linguistic functions, such as speaking style or speech emotions, does not benefit from the same amount of attention from the linguistic community. Meanwhile, speech engineers have realized spectacular advances in the past decade creating extremely realistic synthetic voices [1] which are now integrated into voice interfaces that are increasingly present in our everyday lives, such as the voice assistants and conversational/virtual agents. However, these intelligible and natural voices still clearly lack expressiveness and adaptability which greatly limits the interaction between humans and machines (see for instance [2]). Expressivity is the next frontier of speech research at the interface of linguistics and technology, as shown by the recent increase of research in this domain [3]. Consequently, there is a clear need to better understand the expressivity of the human voice which is mainly conveyed through speech prosody. In particular, there is a growing interest of the engineering community in the statistical modelling of prosody [4, 5, 6, 7, 8, 9, 10], expressive speech synthesis [11, 12, 13], and conversion of neutral to expressive speech [14, 15, 16].

Nowadays, speech expressivity is generally equated to speech emotions though the scope of expressivity includes but is not restricted to the primary emotions as denoted by Ekman [17]. This limitation is probably due to the difficulty of converging to an agreement on the terminology used to describe the various and subtle forms of expressivity in speech. Accordingly, the study of speech expressivity is generally limited to dedicated speech emotion databases as interpreted by actors or to audio books read by professional readers (mostly in English, and sometimes in French or German [18]). In the past decade, speech emotion research has mainly focused on acted emotional speech: from its original form in which an actor is asked to interpret a short text with a given emotion [19] to more open and spontaneous forms in which two actors freely improvise based on a given scenario and then asked to rate their own speech emotions with categories or on valence/arousal continuous scales [20, 21]. Moreover, these databases have been created only for the purpose of speech emotion recognition, but not for the modelling of speech prosody and neither for speech synthesis and conversion.

However, speech expressivity is not restricted to primary emotions. For instance, to deal with more subtle variation, the circumplex model, in which emotions are categorized in a bidimensional space, was proposed by psychologists [22]. Attitude was firstly equated to the first dimension (valence) of this model [23]. A distinction between emotion and attitude was done in [24] by defining emotion as a speaker state and attitude as a kind of behaviour. This distinction was later refined in [25], by defining attitude as a predictor of social behaviour. The attitudinal aspect of expressiveness of course differs from the primary emotions, and create a distinction between the propositional attitude (towards an utterance: irony, doubt, etc... [26]) and the social attitudes (towards a person: dominant, friendly, seductive, distant, etc...). This last dimension has been recently investigated in the study of the role of speech prosody in neurosciences [27]. Moreover, all of the existing speech databases miss an essential aspect of speech prosody: its *variety* [28]. There is always only one realization of each utterance, while any utterance can be obviously realized with many prosodies, some being functionally equivalent, some having various degrees of expressivity. There is a clear need for speech database that would tackle these issues and allow a diversification of the research in speech prosody and expressivity, and with sufficient data to allow learning generative models.

This paper presents Att-HACK, the first large database of acted speech with social attitudes. This database comprises the recordings of 25 speakers (M/F) interpreting 100 utterances in 4 different social attitudes: friendly, seductive, dominant, distant. In a given attitude, each utterance is repeated differently between 3 and 5 times in order to provide access to the inherent prosodic variety of speech. This represents a total of about 30 hours of speech, which comes with audio signals, orthographic text transcription, F0 analysis, and phonetic alignments. The Att-HACK * is freely available for research under a Creative

*. http://www.openslr.org/88/

10.21437/SpeechProsody.2020-152

Commons (BY/NC/ND) Licence. The remaining of the paper is organized as follows: in Section 2 we describe the conception of Att-HACK. A full description of the database is proposed in Section 3, with some illustrations of the F0 obtained in the 4 social attitudes by some of the speakers.

2. Conception of Att-HACK

In this Section, the full conception of the Att-HACK speech database of social attitudes is described, from the initial choice of social attitudes as a continuation of speech expressivity research and including the details of the speech database design.

2.1. Models of Emotions

The means by which humans construct representations of emotions is still an open and debated issue in the field of psychology and affective sciences: today, two fundamental models are used for the description of emotions. The first approach is categorial and supports the idea that emotions are discrete and relate to concepts that are essentially distinct. A discrete emotion theory was developed by Ekman et al. in 1992 [17] in which six basic emotions were proposed: anger, disgust, fear, happiness, sadness, and surprise. The second approach is dimensional and supports that emotions are defined according to at least one dimension. In the circumplex model [22], emotion is defined on two dimensions: valence (positive–negative) and arousal (passive-active). A third dimension denoted dominance (submissiveness-dominance) was added to compose the PAD (Pleasure, Arousal, Dominance) three-factors model for speech emotions [29].

2.2. Definition and Choice of Social Attitudes

The choice of social attitudes is methodologically rooted with the idea of extending the study of expressivity from primary emotions to more subtle forms. As initially inspired by [25], social attitudes allows to represent the attitude of a speaker towards its interlocutor during an interaction. Such social attitudes differ from emotions which are internal state of a speaker and from propositionnal attitudes which are the attitude of a speaker towards an utterance.

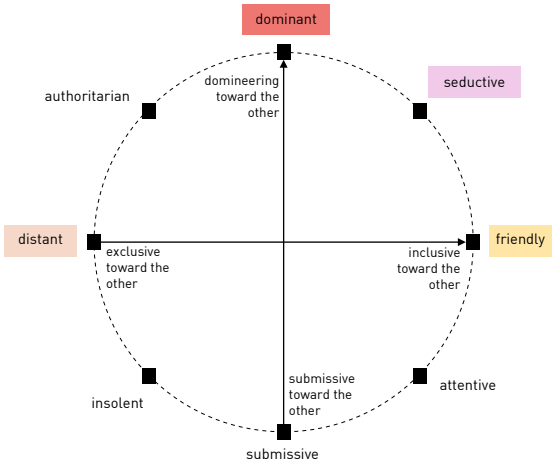

FIGURE 1 – Social attitudes represented in a 2-D space inspired by Jean Julien Aucouturier's categorization for musicians playing attitudes [30].

Following this initial definition, Jean-Julien Aucouturier et al. [30] recently proposed an original categorization for musicians' playing attitudes inspired by Leary's rose model [31]. The attitudes are described in a bidimensional space, the first dimension reflects the hostility/friendliness (alternatively, negative/positive) towards the other musician while the second dimension reflect the position of the musician in a social hierarchy (subordination/dominance). Our proposed choice of social attitudes is in the continuation of these two precursory works. In this paper, four social attitudes were defined by sampling the valence and dominance dimensions during a speech interaction: friendly, seductive, dominant and distant (figure 1). This includes one negative (distant) and three positives (friendly, seductive, dominant) attitudes sampled in the semi-space of neutral to high-hierarchy in the dominance space.

The four attitudes are described as follows :

— **friendly**: you are pleasant and benevolent, you care about others' preferences, you act towards the others independently from your own situation.

— **seductive**: everything in your behaviour aims at charming the others, to make them love you, you do not care about others' preferences but you are ready for anything to seduce them even if you have to fake benevolence.

— **dominant**: you are self confident, sure of your own superiority, you do not care about others' preferences, everything in your behaviour is dedicated to make the others obey and listen to you without imposing anything explicitly.

— **distant**: you (barely) do not care about the others, you are uncommunicative, you do not care about others' preferences.

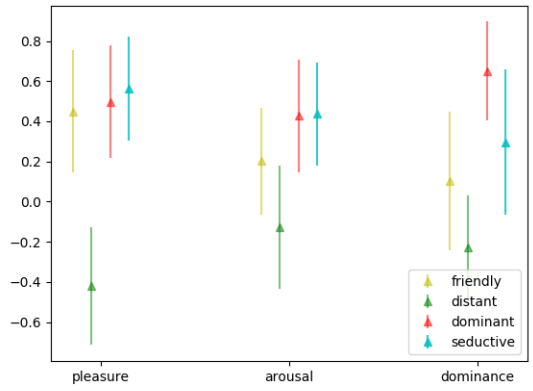

FIGURE 2 – Social attitudes represented with the PAD model proposed by Russel & Mehrabian [29]

An attempt to describe our four attitudes through using the Russell's PAD model has been initiated by recomposing those attitudes with some of the 151 emotion-denoting terms subjects were asked to rate in [29] using valence, arousal, and dominance. Let us assume that our friendly is a mix of {friendly, humble, curious, respectful, kind}, *seductive* a mix of {affectionate, sexually excited, controlling}, *dominant* a mix of {influential, domineering, bold} and *distant* a mix of {timid,

bored, inhibited, uninterested, detached, shy, snobbish lonely}. The figure 2 shows our four social attitudes categorized according to the three Russel's dimensions (by averaging of the results obtained by Russel for the considered groups).

2.3. Set of sentences

The set of sentences used for the creation of the database has been designed in French as inspired by the corpus of propositionnal attitudes proposed by Morlec in [32]. The proposed sentences have been designed according to the following criteria: 1) to avoid introducing a bias due to the expressive content of the text, the sentences were designed to be as neutral as possible ; 2) to reduce the prosodic variability due to the text structure, the sentences were designed with a limited and controlled linguistic complexity, by using simple syntactic structures and short sentences ; 3) Finally, the sentences were designed in a way they remain plausible in each social attitude.

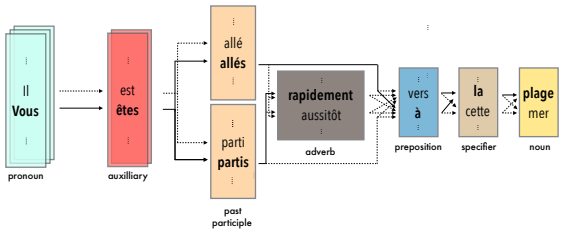

FIGURE 3 – Phrase generator functioning for the above quoted phrases, chosen words are in bold, dotted lines represent possible choices for the algorithm

Accordingly, we constructed a set of 100 sentences (from 2 to 8 syllables) corresponding to simple everyday life situations (classic situations in classic socialization places like home, restaurant, workplace, ...). For this purpose, we designed a phrase generator that builds phrases from semantic nucleus ({pronoun/noun + verb} or {pronoun/noun + auxilliary}) by randomly picking words in dictionaries in order to guarantee the phrases are always conceived the same way. We randomly kept 100 phrases among the 10000 generated ones to build our set of sentences, a sample of 10 sentences is listed below.

Oui
Bonjour
C'est vrai
A demain Paul
Bonne journée Marie
Il est tard à Londres
Vous êtes allés à la plage
Vous êtes partis rapidement
Impossible, attendons un peu
C'est vrai, allons prendre un café

Figure 3 depicts the functioning of the automatic sentence generator considering the example of two sentences created from the nucleus "Vous êtes ..." ("You are ...").

3. Description of Att-HACK

3.1. Audio recordings

To feed this database, twenty-five actors were recorded in professional studios at Ircam. It consisted of 4 hours sessions during which one actor had to play 100 sentences in the four different social attitudes, proposing from 3 to 6 different versions of each sentence in each attitude. At the beginning of each session, the four attitudes were shortly described as stipulated above. Those descriptions have been used as acting options, actors were told to act the way they felt regarding each attitude denomination and not necessary in regards to those descriptions. The actors were told to be as natural as possible, no other information was given during the session.

For the recording, we used a Neumann U87 static microphone plugged into a RME fireface interface synchronized with the ProTools software. All audio recordings were made with a sampling rate frequency of 44.1 kHz and quantization of 16 bits per sample. The recording were made in two different studios depending on their availability. A patch implemented with the Max for Live software † was used to provide a visual interface to the actor, displaying on a screen in front of the actor the sentence to be read and the expected attitude, this display being monitored by a sound engineer. The patch also allowed to store the time codes corresponding to each sentence, which were used after the session to segment and name automatically the continuous recording made with the ProTools software. At the end of a session, we had at least 2500 audio files for each actor. This amount of recordings has been manually sorted to remove corrupted files and utterances that were judged as being not natural enough or badly acted. At the time of writing, the Att-HACK database is composed of more than 22,000 expressive utterances, which will be completed in the months to come to reach around 50,000 utterances and 30 hours of speech.

3.2. Speech to text alignment

Speech-to-text alignment was performed by using the ircamAlign software [33] which is based on the HTK toolbox and the Lia_phon French phonetizer and learned on the BREF French database of read speech. Phonetic alignment is first processed by ircamAlign and followed by rule-based syllable segmentation and a simple phrase segmentation. The resulting alignment are stored in dedicated lab files, a text format indicating at each line the starting time, ending time, and label corresponding to the text sentence and audio recording.

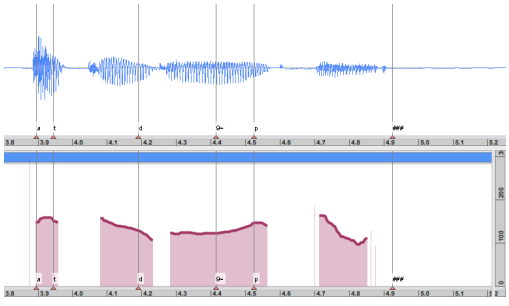

FIGURE 4 – Syllable segmentation (above) and F0 values as estimated on voiced frames (below)

†. https://www.ableton.com/en/live/max-for-live/

3.3. F0 estimation

The fundamental frequency of the speakers was estimated by using the SWIPEP algorithm [34] with a minimum pitch value of 75 Hz, a maximum pitch value of 450 Hz, and a hop size of 5 ms, without any post-processing for correcting or smoothing the raw pitch values. The voiced/unvoiced decision was computed from the pitch strength associated with the pitch value estimate with a threshold of 0.25 (the pitch strength being a value between 0 and 1 corresponding to the periodicity of the speech frame).

Figure 4 presents the phonetic segmentation of the speech waveform (above) in syllables and the corresponding F0 values (below) for the sentence *"Attendons un peu"* using the Audiosculpt software ‡.

4. Preliminary investigation

A preliminary investigation was conducted to illustrate the content of the Att-HACK speech database of social attitudes.

4.1. Extraction of F0 contours

A F0 contour was extracted for each syllable in order to illustrate the F0 patterns realized by the actors for the different social attitudes. The F0 contour of a syllable was identified as the one corresponding the longest sequence of F0 values considered as voiced over the syllable, i.e. the longest F0 segment for which each F0 value corresponds to a pitch strength value which is above a given threshold (in this paper, 0.3).

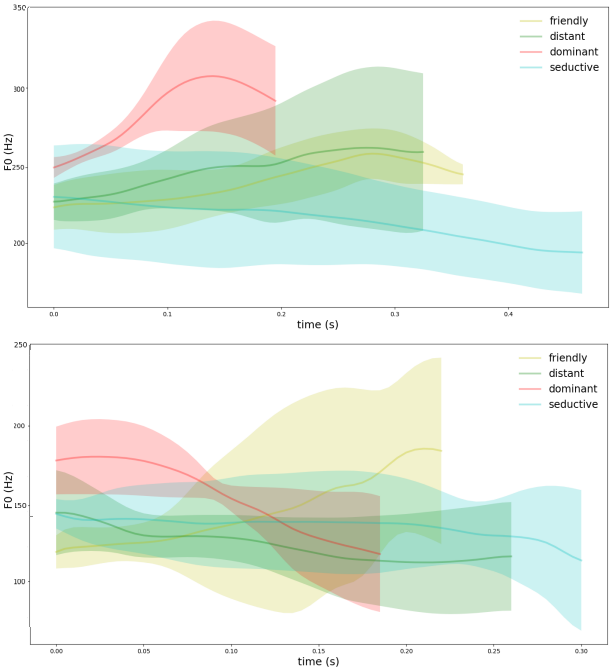

FIGURE 5 – F0 contours mean (solid line) and standard deviation (filled with color) for the phrase *"Oui"* for a female speaker (above) and a male speaker (below)

‡. https://forum.ircam.fr/projects/detail/audiosculpt/

gender	Friendly		Seductive		Dominant		Distant	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
female	208	11	186	12	207	10	186	10
male	112	8	113	10	119	8	104	7
global	160	10	150	11	163	9	145	9

TABLE 1 – Syllable F0 contours means and standard deviations (in Hz) for female and male speakers of Att-HACK

gender	Friendly		Seductive		Dominant		Distant	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
female	404	151	426	141	399	143	440	169
male	410	133	440	145	413	137	471	177
global	407	144	431	142	405	141	452	172

TABLE 2 – Syllable durations means and standard deviations (in ms) for female and male speakers of Att-HACK

4.2. F0 Statistics vs Attitudes

A preliminary investigation was conducted to compare the F0 statistics of the actors across the social attitudes. We computed mean and standard deviation statistics on F0 segments extracted as stipulated above, for each speaker and attitude. These statistics are reported in Table 2, including global statistics (female, male and mixed-gender). It is to be noted that only the first part of the recordings has yet been post processed at the time of writing.

Figure 5 illustrates F0 pattern distributions obtained for a given sentence with the four attitudes, each attitude being represented by a dedicated color. In each color, the solid line represents the F0 pattern obtained by averaging the variations realized by the actor, the area filled with color represents the corresponding standard deviation, and the length of the pattern the corresponding mean duration. This illustration reveals that distinctive F0 patterns are associated with the social attitudes, and also highlights the diversity of strategies employed by actors to communicate a social attitude.

5. Conclusion

This paper presents Att-HACK, a first attempt to widen the scope of expressivity in speech, by providing a database of acted speech with social attitudes: friendly, seductive, dominant, and distant. The proposed database comprises 25 speakers interpreting 100 utterances in 4 social attitudes, with 3-5 repetitions each per attitude proving a great prosodic variety for a total of around 28 hours of expressive speech. The Att-HACK is freely available for academic research under Creative Commons Licence.

6. Acknowledgements

This research has been supported by the French Ph2D/IDF MoVE project on MOdelling of speech attitudes and application to an expressive conversationnal agent and funded by the Ile-De-France region and by the French ANR project TheVoice: ANR-17-CE23-0025.

7. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," 2017.

[2] C. P. P. W. M. G. Castellano, M. Mancini, "Expressive copying behavior for social agents: a perceptual analysis," *IEEE Trans Syst, Man Cybern*, vol. 42, no. 3, 2012.

[3] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," 2018.

[4] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis," in *Interspeech*, Brisbane, Australia, 2008, pp. 2274–2277.

[5] N. Obin, A. Lacheret, and X. Rodet, "Stylization and Trajectory Modelling of Short and Long Term Speech Prosody Variations," in *Interspeech*, Florence, Italy, 2011, pp. 2029–2032.

[6] X. Yin, M. Lei, Y. Qian, F. K. Soong, L. He, Z.-H. Ling, and L.-R. Dai, "Modeling F0 trajectories in hierarchically structured deep neural networks," *Speech Communication*, vol. 76, pp. 82–92, 2016.

[7] X. Wang, S. Takaki, and J. Yamagishi, "An RNN-Based Quantized F0 Model with Multi-Tier Feedback Links for Text-to-Speech Synthesis," in *Interspeech*, 2017, pp. 1059–1063.

[8] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with arbitrary scales F0 based on wavelet transform," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, 2017.

[9] N. Obin and J. Belião, "Sparse coding of pitch contours with deep auto-encoders," in *International Conference on Speech Prosody*, 2018, pp. 799–803.

[10] G. Branislav, G. Bailly, O. Mohammed, Y. Xu, and P. N. Garner, "A variational prosody model for the decomposition and synthesis of speech prosody," in *ArXiv e-prints*, 2018. [Online]. Available: arxiv.org/abs/1806.08685

[11] N. Obin, "MeLos: Analysis and Modelling of Speech Prosody and Speaking Style," PhD. Thesis, IRCAM - UPMC, 2011.

[12] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive tts," 03 2012.

[13] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder," *arXiv e-prints*, p. arXiv:1804.02135, Apr 2018.

[14] C. Veaux and X. Rodet, "Intonation conversion from neutral to expressive speech," in *Interspeech*, Jan. 2011, pp. 2765–2768.

[15] C. Robinson, N. Obin, and A. Roebel, "Sequence-to-sequence modelling of f0 for speech emotion conversion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6830–6834.

[16] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[17] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992. [Online]. Available: https://doi.org/10.1080/02699939208411068

[18] L. Chen, N. Braunschweiler, and M. Gales, "Speaker and expression factorization for audiobook data: Expressiveness and transplantation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, pp. 605–618, 04 2015.

[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Interspeech*, Lisbon, Portugal, 2005.

[20] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] G. McKeown, M. Valstar, M. Pantic, and R. Cowie, "The SEMAINE Corpus of Emotionally Coloured Character Interactions," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2010.

[22] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 12 1980.

[23] I. Ajzen and M. Fishbein, *Understanding attitudes and predicting social behaviour*. Englewood Cliffs, New Jersey: Prentice-Hall, 1980.

[24] E. Couper-Kuhlen, *An Introduction to English Prosody*, 1986.

[25] A. Wichmann, "The attitudinal effects of prosody, and how they relate to emotion," in *ITRW on Speech and Prosody*, Newcastle, UK, 2000.

[26] G. Leech, *Principles of Pragmatics*. London: Sage Longman, 1983.

[27] 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*, pp. 3972–3977, 2018.

[28] N. Obin, C. Veaux, and P. Lanchantin, "Making Sense of Variations: Introducing Alternatives in Speech Synthesis," in *Speech Prosody*, Shanghai, China, 2012.

[29] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, pp. 273–294, 09 1977.

[30] J.-J. Aucouturier and C. Canonne, "Musical friends and foes: The social cognition of affiliation and control in improvised interactions," *Cognition*, vol. 161, pp. 94–108, 04 2017.

[31] T. Leary, *Interpersonal diagnosis of personality*, 1957.

[32] Y. Morlec, "Génération multiparamétrique de la prosodie du français par apprentissage automatique," PhD. Thesis, Institut de la Communication Parlée, Grenoble, 1997.

[33] P. Lanchantin, A. Morris, X. Rodet, and C. Veaux, "Automatic Phoneme Segmentation with Relaxed Textual Constraints," in *International Conference on Language Resources and Evaluation*, Marrakech, Morroco, 2008, pp. 2403–2407.

[34] A. Camacho, "SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music," PhD. Thesis, University of Florida, 2007. [Online]. Available: http://www.cise.ufl.edu/~acamacho