INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Affective attributes of French caregivers' professional speech

Jean-Luc Rouas[1], *Yaru Wu*[2,3,4], *Takaaki Shochi*[1,5]

[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
[2]CRISCO/UR4255, Université de Caen Normandie, 14000 Caen, France
[3]Laboratoire de Phonétique et Phonologie UMR7018, CNRS-Sorbonne Nouvelle, France
[4]LISN UMR 9015, CNRS, Univ. Paris-Saclay, France
[5]CLLE CNRS UMR 5263, Bordeaux, France

rouas@labri.fr, yaru.wu@unicaen.fr, takaaki.shochi@labri.fr

Abstract

In this paper, we detail our approach to studying the vocal characteristics of caregivers in French retirement homes. To achieve this goal, we conducted recordings of 20 professional caregivers across two retirement homes. Using headset microphones connected to smartphones, we were able to capture the caregivers' speech while allowing them complete freedom of movement without compromising sound quality. The recordings consisted of three tasks: reading text, informal interviews, and professional role-play scenarios with a fictitious patient. We processed the recordings using an automatic speech recognition system, which provided word or phone sequences and their corresponding timestamps. Our analysis focused on identifying differences in emotional tone, lexical content, speech rate, fundamental frequency, and intensity between spontaneous speech conditions. Ultimately, our aim is to develop automated training tools that capture the unique vocal characteristics of professional caregivers.

Index Terms: Affective speech, caregivers voice, speaking styles, spontaneous speech

1. Introduction

Social interaction among humans involves the exchange of social information through various modalities, such as voice, eye contact, gestures, and facial expressions [1, 2]. The prosodic features of affective expressiveness are a crucial modality that conveys a speaker's various affective meanings [3, 4].

Verbal communication is crucial for interactions with dependent elderly individuals, but in hospital settings, caregivers are often occupied with their work, resulting in brief interactions with patients. A report indicated that verbal communication lasted only two minutes per day for a bedridden dementia patient in a long-term care facility [5]. Caregivers may feel disheartened by patients who are uncommunicative and provide irrelevant responses.

Studies have demonstrated that intentional positive expressiveness displayed by professional caregivers can have a significant impact on elderly dementia patients [6]. For example, [7] investigated the effect of vocally expressed emotions and moods by professional caregivers on individuals with severe dementia. The results indicated that caregiver singing had a positive effect on patients, enhancing their positive emotions and reducing their aggressive behavior.

The "Humanitude" method is designed to promote positive interactions through the development of effective communication skills, which are based on face-to-face interactions, verbal communication, and touch interaction. Numerous studies have demonstrated that this method results in a significant reduction (88.5%) in patients' aggressive behavior and a decreased need for neuroleptic medication [8, 9, 10].

Regarding vocal communication skills, the "Humanitude" method relies on phonetic and lexicological elements, as well as a technique called *auto-feedback*, where caregivers continue speaking without interruption, even when care recipients provide inadequate responses. Therefore, two categories of parameters need to be studied: prosodic parameters (intensity, rate, melody) that should align with the recommended soft, calm, and melodious voice, and lexical elements that aim to convey positive emotions.

The present study aims to investigate the similarities and differences between three types of speech styles, namely 1) text reading, 2) spontaneous talk, and 3) professional talk. In this paper, we aim to identify the emotional content of the discourse using both acoustic and lexical cues.

This paper briefly summarizes in Section 2 the recording protocol of the "tender care" corpus produced by French professional caregivers with the description of the specific equipment used to allow freedom of movement, the tasks carried out and the recording settings. Then, Section 3 describes how we preprocessed the files to obtain the phonetic transcription which is then used to extract acoustic features on Inter-Pausal Units. The analysis of the features is carried out in Section 4 and we discuss the results in section 5.

2. Recording protocol

2.1. Equipment

To ensure complete mobility of the recorded subjects, we designed a fully autonomous recording setup. We outfitted our subjects with a high-quality directional microphone, the DPA 4288 CORE headset, which was connected to an iRIG PRO preamplifier that, in turn, was connected to a Samsung Galaxy A51 smartphone. Both the preamplifier and the smartphone were placed in a waist bag, which provided complete freedom of movement, while the headset microphone ensured high-quality recording.

2.2. Tasks

2.2.1. Text Reading

The first task involves reading a text, specifically *The North Wind and the Sun*. This particular text has been used for over a century by the International Phonetic Association to illustrate a variety of dialects and languages from around the world. For our French-speaking participants, the text is presented in its French version, *La bise et le soleil*. The purpose of this task is to record a controlled voice sample in a highly structured context. Typically, the text reading takes less than a minute.

10.21437/Interspeech.2023-848

2.2.2. *Informal interview*

The interview task involves a questionnaire comprising open-ended questions that focus on the caregiver's work and their daily routine. The purpose of this task is to encourage the caregiver to talk about themselves as much as possible. To avoid making the interviewer feel embarrassed, impersonal questions about work were selected. This exercise also helps to build the caregiver's confidence.

2.2.3. *Professional care task*

The aim of the experiment was to record the speech of caregivers while performing a care task on a fictitious unresponsive patient. The care task chosen was dressing up, which involved buttoning a shirt and techniques for waking up the body. Following the dressing, the caregiver uprighted the fictitious patient and helped them walk. To evaluate the technique of *autofeedback*, the fictitious patient remained completely mute and allowed the caregiver to provide care. The context was designed to be familiar to the caregiver, with a bed similar to those used for regular patients and partitions creating intimacy with the fictitious patient.

2.3. Collected data

The audio recordings were acquired in two retirement homes for dependent elderly people located in southwestern France: *Les Balcons du Lot* in Prayssac and *Les résidences du Quercy Blanc* in Castelnau-Montratier. Three recording sessions were conducted: two at the Prayssac establishment on September 24, 2021, and March 25, 2022, and one at Castelnau-Montratier on November 24, 2021.

A total of 26 participants were recorded during the three sessions, of which 21 were female and 4 were male. For the analysis, we excluded the 4 male participants and 1 recording of a female participant due to poor recording quality. The total duration of the 20 remaining participants' recordings is 2 hours and 30 minutes. The duration of each task and the mean duration per speaker per task are given in Table 1.

Task	mean dur.	total dur.
Text reading	45.0 s.	15 m. 03 s.
Interview	134.8 s.	44 m. 56 s.
Professional care	188.9 s.	62 m. 58 s.

Table 1: *Mean duration per speaker per task and total recorded duration per task. All 20 recorded subjects participated in each task.*

3. Features

3.1. Automatic orthographic and phonetic transcription

To obtain orthographic and phonetic transcriptions, an automatic system was used based on the Kaldi framework [11], which was trained using the ESTER database [12]. The system used a *Time Delay Neural Network* (TDNN) coupled with a hidden Markov model. The TDNN had 7 layers, with 1024 units in each layer. The input for the acoustic model was a 40-dimensional high-resolution MFCC vector concatenated with a 100-dimensional I-vector [13].

3.2. Parameters extraction

Automatic segmentation in Inter-Pausal Units (IPUs) was carried out using the output of the automatic transcription system with a threshold of 250 ms. Each IPU was further divided into smaller segments between pauses. Using the snack routines [14], fundamental frequency (F0) and intensity were extracted at 10 ms intervals. The mean and standard deviation of the F0 and intensity values were calculated for each IPU. To facilitate analysis, the fundamental frequency was converted to semitones (ST) relative to a reference value (50 Hz).

The speech rate was determined for each IPU by analyzing the phonetic transcription system and calculating the number of phones per second.

3.3. Arousal and valence detection

Dimensional emotion recognition was carried out using the *SpeechDimEmo* software [1]. This software allowed us to retrieve continuous values of valence and arousal using models based on the Recola dataset [15]. The Recola corpus is widely recognized as a benchmark for evaluating the performance of emotion recognition systems. It comprises recordings of natural interactions among French-speaking individuals in laboratory settings.

While arousal refers to the level of physiological activation or intensity of an emotion, ranging from low arousal (e.g., feeling calm or relaxed) to high arousal (e.g., feeling excited or anxious), valence, on the other hand, refers to the positive or negative quality of an emotion, ranging from negative valence (e.g., feeling sad or angry) to positive valence (e.g., feeling happy or content).

In this study, we used the setting with 80-dimension log Mel filterbank (MFB) features and a 1-layer GRU model with a hidden layer of dimension 32 as in [16].

As we do not expect strong emotional content in our data, we are looking forward to find low arousal values for all tasks in this study. Valence is however expected to be higher for the professional care task in which our speakers are supposed to express positivity.

3.4. Lexical content analysis

The lexical content analysis was performed using the EMOTAIX-Tropes text analysis software [17]. More precisely, EMOTAIX is a scenario exploitable with the Tropes software that allows for the quantification of emotional lexicons present in textual corpora. It is organized into 56 hierarchical semantic categories. At the highest level, 5 major categories are proposed: emotions are classified according to their valence into two categories: positive or negative. In addition to these two super-levels, there are levels of surprise, impassivity, and unspecified emotions. Negative and positive emotions are then broken down into 3 lower and increasingly specific levels. The EMOTAIX-Tropes software has been recently used to analyse self-defining autobiographic memories of French speakers in [18] and emotional qualifications of tweets during the covid-19 confinement in [19].

In order to perform the lexical content analysis, we aggregated all texts resulting from the automatic speech recognition system for each speaking style (interview and professional care) before analysis. This aggregation leads to a total number of 5640 words for the interview condition and 5309 words for the professional care condition.

[1] `https://github.com/SinaAlisamir/SpeechDimEmo`

As for the dimensional emotional values, we are expecting to find more positive words in the discourse transcribed for the professional care task than for other conditions.

4. Results

4.1. Acoustic parameters

To explore the global correlation between the acoustic features of F0 (mean, standard deviation converted to semitones), intensity (mean, standard deviation), speech rate (phone rate, number of IPUs) and duration measurements (mean phone duration and duration of IPUs), a Principal Component Analysis (PCA) was carried out (Figure 1) using FactoMineR and factoextra packages in R [20]. Before computing the PCA, all acoustic values were converted into z-scores setting average value of reading dataset as reference value for each parameter. In addition, linear mixed models (LMM) were also used to analyze arousal, valence and acoustic features using the *lme4* package in R. A model is conducted for each parameter. We included speech style as a fixed effect for all models and intercepts were included for subject and item.

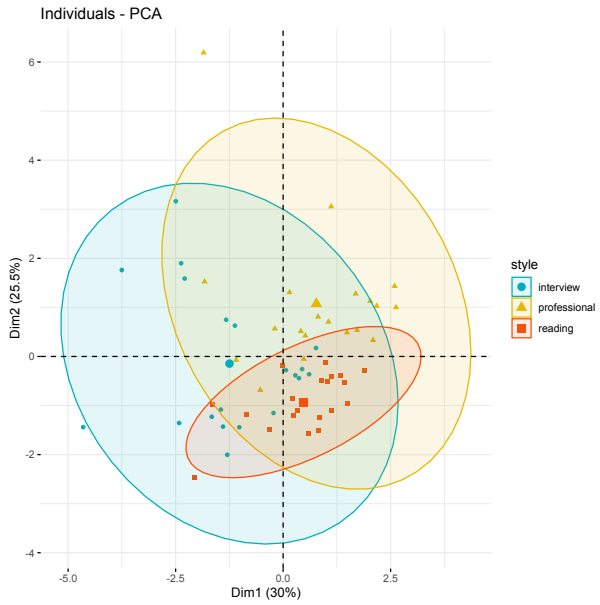

Figure 1: *Distribution of 20 subjects' attributions categorized by three different speaking styles. Each ellipse is fixed at 0.95 confidence level.*

According to Figure 1, we observe a global difference of three different speaking styles. This acoustic difference of these three speaking styles was significant and confirmed in [21]. The reading task leads to acoustically similar features and few variations compared to other speaking styles. it is probably due to this emotionally neutral speaking style. Even if the two other speaking styles have also some common acoustic features with reading, spontaneous and professional care voices showed an important acoustic variation.

Figure 2 highlights the important contribution of all acoustic parameters (i.e. F0, duration and intensity variables) in order to determine the principal components of all speaking styles on first two dimensions. According to Figure 2, F0 and duration features showed an important contribution to change various speaking style. More concretely, fundamental frequency is

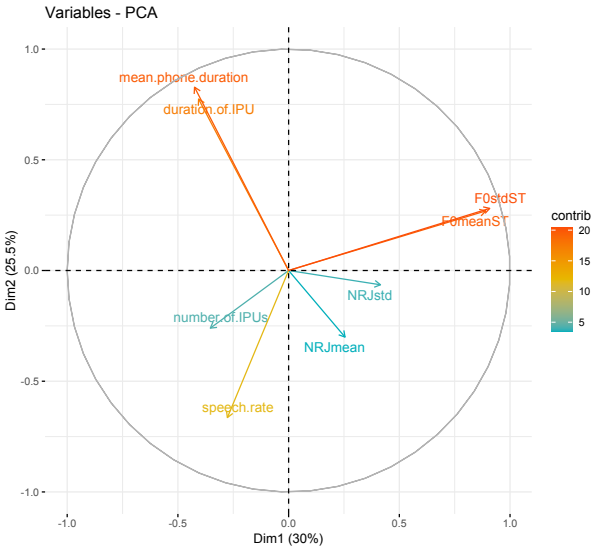

Figure 2: *Degrees of contribution of 8 acoustic parameters of three speaking styles computed by Principal Component Analysis.*

on average lower for the informal interview than for the professional voice. Therefore, this result indicates that caregivers tend to use intentionally higher voice in the professional situation.

With regard to the duration, Linear Mixed models showed that our speakers tend to speak faster when reading [β = 14.48; t = 6.99; SE = 2.07] or being interviewed [β = 11.17; t = 5.62; SE = 1.99] than when they are caring. This phenomenon is also expected since when trying a keep the steady vocal flow when caring, our caregivers naturally tend to slow their speech rate. A lower speech rate is also an indication of a calmer voice.

Concerning intensity, this analysis did not show an important contribution to the three speaking styles.

4.2. Arousal and valence detection

Figure 3 presents the arousal values as a function of the three speech styles (from left to right : reading, interview, professional care). The results show that the arousal values are lower for reading tasks [β = -0.019; t = -2.859; SE = 0.006] than for professional voice. No significant difference was observed between interview and professional care, as far as arousal values are concerned. A low arousal value is indeed expected for the reading task. The slightly higher arousal values observed for the two other tasks may be linked to a higher activation required by spontaneous speech.

Figure 4 shows the valence values as a function of the three speech styles. The results show that the valence values are lower for the reading tasks [β = -0.057; t = -3.666; SE = 0.016] and interviews [β = -0.070; t = -4.480; SE = 0.016] than for professional voice.

Valence values reflect our expectations as caregivers should carry positive emotions when caring. The text reading task and the spontaneous speech interview both lead to lower valence values characterising a more neutral voice.

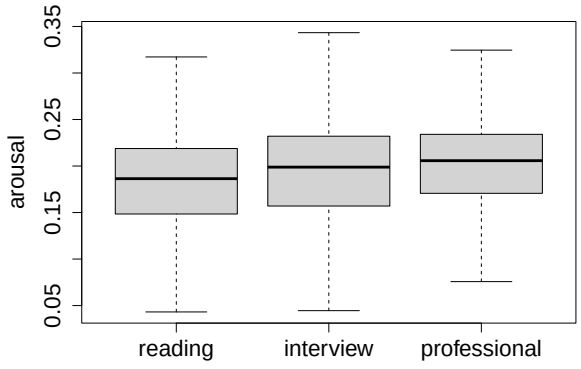

Figure 3: *Boxplot of arousal values average over IPUs for the three speaking styles.*

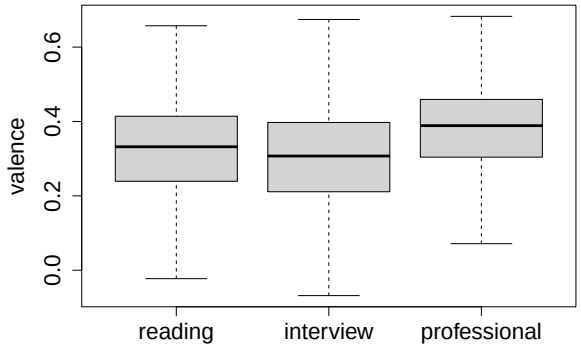

Figure 4: *Boxplot of valence values average over IPUs for the three speaking styles.*

4.3. Lexical analysis

The table 2 shows the results of an analysis of emotional words found by the EMOTAIX software for two different speaking styles: interview and professional care. The software identified a total of 117 emotional words in the interview style and 227 emotional words in the professional care style.

The table also shows the distribution of emotions for each speaking style. For the interview style, 14% of the emotional words were unspecified, 34% were negative, and 52% were positive. For the professional care style, 19% of the emotional words were unspecified, 16% were negative, and 61% were positive.

Words with unspecified emotion are polysemic words for which the emotional meaning is not clear, i.e. words that carry not only a literal meaning and a figurative meaning, but also several literal meanings. A contextual analysis is thus needed to correctly classify those words. As our objective here is to provide an automatic analysis of the discourse, we have decided to not consider words with unspecified emotions.

Table 2: *number of emotional words found by the EMOTAIX software and their proportion in unspecified, positive and negative emotions for the interview and the professional care speaking styles. Total number of words : 5640 (interview), 5309 (professional).*

Emotional words	Interview (n=117)	Professional care (n=227)
Unspecified emotions	14%	19%
Negative emotions	34%	16%
Positive emotions	52%	61%

This analysis suggests that emotional words can be found in both speaking styles. More emotional words are used in professional care (4.3% of total words) than in the interview task (2%). The results also indicate that the professional care style uses a slightly higher proportion of word conveying positive emotions and a lower proportion of words conveying negative emotions compared to the interview style.

5. Conclusion

To summarize, our analysis focused on the speaking styles of caregivers in three different conditions: reading tasks, interviews, and professional care. The results of our acoustic analysis indicate that, during professional care, caregivers tend to use longer sentences, slower speech rates, and higher pitches with more melodic variations compared to other speaking styles. However, intensity did not show any important contribution in the three speaking styles.

Our analysis of emotional dimensions aligns with our expectations, with no particular differences in arousal and a more positive valence observed in the professional care condition. These findings are further supported by our lexical analysis, which indicates a higher frequency of positive words used during professional care compared to interviews.

These results are consistent with the use of soft, calm, and melodious voices, as well as the use of positive language aimed at conveying positive emotions during the application of the ”humanitude” technique in professional care.

6. Acknowledgements

The authors thanks the founders of ”Humanitude” Yves Gineste and Rosette Marescotti for their help in building this project. Many thanks to Jean-Yves Nou, Hervé Tomassi and especially Aurélie Rives for allowing us to record in professional settings. We are deeply indebted to all the professional caregivers we recorded for their trust and their time. This research was supported by the French RNMSH “Humavox” project.

7. References

[1] A. Wichmann, “The attitunal effects of prosody, and how they relate to emotion,” in *Proc. of ISCA Workshop on Speech and Emotion*, Newcastle, 2000, pp. 143–148.

[2] N. Campbell, “Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language,” *Language Resources and Evaluation*, vol. 39, no. 1, pp. 109–118, Feb. 2005.

[3] K. R. Scherer and T. Bänziger, “Emotional expression in prosody: A review and an agenda for future research,” in *Speech Prosody*, Nara, Japan, 2004.

[4] A. Rilliard, T. Shochi, J.-C. Martin, D. Erickson, and V. Aubergé, "Multimodal indices to Japanese and French prosodically expressed social affects," *Language and speech*, vol. 52, no. 2-3, pp. 223–243, 2009.

[5] Y. Gineste, R. Marescotti, and J. Pellissier, "L'humanitude dans les soins," *Recherche en soins infirmiers*, vol. 94, no. 3, pp. 42–55, 2008.

[6] Y. Gineste and J. Pellissier, *Humanitude*, nouvelle édition: armand colin ed., 2007.

[7] E. Götell, S. Brown, and S.-L. Ekman, "The influence of caregiver singing and background music on vocally expressed emotions and moods in dementia care: A qualitative analysis," *International Journal of Nursing Studies*, vol. 46, no. 4, pp. 422–430, Apr. 2009.

[8] M. Honda, M. Mori, S. Hayashi, K. Moriya, R. Marescotti, and Y. Gineste, "The effectiveness of French origin dementia care method; Humanitude to acute care hospitals in Japan," *European Geriatric Medicine*, vol. 4, p. S207, Sep. 2013.

[9] M. Honda, M. Ito, S. Ishikawa, Y. Takebayashi, and L. Tierney, "Reduction of Behavioral Psychological Symptoms of Dementia by Multimodal Comprehensive Care for Vulnerable Geriatric Patients in an Acute Care Hospital: A Case Series," *Case Reports in Medicine*, vol. 2016, p. 4813196, 2016.

[10] M. Ito and M. Honda, "An examination of the influence of Humanitude caregiving on the behavior of older adults with dementia in Japan," in *Proceedings of the 8th International Association of Gerontology and Geriatrics European Region Congress*, vol. 2018, 2015.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, Dec. 2011.

[12] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts," in *In In: Proceddings of Interspeech, Brighton (United Kingdom*, 2009.

[13] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2014.

[14] K. Sjölander. (2004) The snack sound toolkit.

[15] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, Apr. 2013, pp. 1–8.

[16] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, B. Lecouteux, F. Portet, S. Rossato, F. Ringeval, D. Schwab, and L. Besacier, "LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 1439–1443.

[17] A. Piolat and R. Bannour, "An example of text analysis software (EMOTAIX-Tropes) use: The influence of anxiety on expressive writing," *Current psychology letters. Behaviour, brain & cognition*, no. Vol. 25, Issue 2, 2009, Apr. 2009.

[18] V. Delvaux, A. Lavallée, F. Degouis, X. Saloppe, J.-L. Nandrino, and T. Pham, "Telling self-defining memories: An acoustic study of natural emotional speech productions," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 1337–1341.

[19] B. Gouvernet, N. Guénolé, P. Chapillon, S. Combaluzier, C. Gouvernet, and T. Plaie, "Impact du 3e confinement lié à la Covid19 sur les émotions des Français : exploration textuelle de 481 601 flux Twitter," *Psychologie Française*, vol. 67, no. 4, pp. 489–507, Dec. 2022.

[20] R Development Core Team, *R: A La nguage and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[21] J.-L. Rouas, Y. Wu, and T. Shochi, "A study on caregivers speech in retirement homes," in *International Congress of Phonetic Sciences (ICPhS)*, Prague, Czech Republic, Aug. 2023.