A STUDY ON CAREGIVERS SPEECH IN RETIREMENT HOMES

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/EA4255, 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-ERSSàB CNRS UMR 5263, Bordeaux
jean-luc.rouas@labri.fr, yaru.wu@unicaen.fr, takaaki.shochi@labri.fr

ABSTRACT

This paper describes our efforts at characterising the caregivers voice in retirement homes. Towards this aim, we recorded 20 professional caregivers in two retirement homes. These caregivers were recorded using a headset microphone connected to a smartphone, allowing a total liberty of movement while keeping a very high sound quality. The caregivers were recorded while performing three different tasks: text reading, informal interview and professional roleplay with a fictive patient. All the recordings were then processed by an automatic speech recognition system that gives words and phones sequences and their timestamps. In our analysis, we focus particularly on the differences between spontaneous speech conditions and look for contrasts in terms of duration, speech rate, fundamental frequency and intensity. Our final aim is to capture the specificity of the professional caregivers voice in order to build automatic training tools.

Keywords: Affective speech, caregivers voice, speaking styles, spontaneous speech

1. INTRODUCTION

Human social interaction is an exchange of social information conveyed by voice, eye contact, gestures and facial expressions [1, 2]. Among these modalities, prosodic features of affective expressiveness are well known as an important modality which conveys a speaker's various affective meanings [3, 4].

Verbal communication is an important part of interaction, especially with dependent elderly people. However, in the hospital setting, most caregivers are focused on their work, and their verbal interactions with patients are brief. One report describes that verbal communication lasted only 2 minutes each day for a bedridden patient with dementia in a long-term care facility [5]. Caregivers are indeed often discouraged by laconic patients who do not respond or give irrelevant answers.

However, it has been shown that an intentional positive expressiveness demonstrated by professional caregiver can have a strong impact in the contexts with elderly dementia patients [6]. For instance, [7] investigated the influence of vocally expressed emotions and moods by professional caregivers to / or with persons with severe dementia. The results showed the positive effect of vocally expressed caregiver singing for these patients. They reported that the singing enhanced patients' positive emotions and diminished their aggressive behavior.

To systematize positive interactions, the "Humanitude" method aims at maintaining good communication skills that must be developed based on face-to-face interactions, verbal communication and touch interaction. Several studies have shown that using this method results in a significant reduction in patients' aggressive behavior and less dependence on neuroleptics [8, 9, 10].

Concerning the vocal communication skills, the "Humanitude" method relies on phonetic and lexicological elements as well as on a technique called *auto-feedback* where caregivers speak without interruption even when care recipients respond inadequately. Thus, two categories of parameters must be studied. On the one hand, prosodic parameters (intensity, speech rate, fundamental frequency) should be in line with the recommended soft, calm and melodious voice, on the other hand, lexical elements should aim to convey positive emotions.

As a first step, this current work investigates common and different points of three speech style text reading, spontaneous talk, professional talk). Then, we aim to identify some prosodic patterns of these speaking styles.

This paper briefly summarizes 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 (Section 2). Then, it describes how we pre-processed the files to obtain the phonetic transcription which is then used to extract acoustic features on Inter-Pausal Units (Section 3), followed by the analysis of the features (Section 4) and the discussion of the results (Section 5).

2. RECORDING PROTOCOL

2.1. Equipment

To ensure total mobility of the recorded participants, we designed a completely autonomous recording setup. We equipped our participants with a headset directional microphone DPA 4288 CORE linked to a iRIG PRO preamplifier connected to a Samsung Galaxy A51 smartphone. Both the preamplifier and the smartphone are fitted in a waist bag ensuring complete freedom of movement while the headset microphone guarantees a high recording quality.

2.2. Tasks

2.2.1. Text Reading

The first task is text reading. The text chosen is *The North Wind and the Sun*. This text has been used for more than a century by the International Phonetic Association to illustrate many of the world's dialects and languages. The text is presented to our French speakers in its French version "La bise et le soleil". The goal here is to record a first control voice in a very structured context. The text reading generally lasts less than a minute.

2.2.2. Informal interview

The interview task consists of a questionnaire with open-ended questions about their work as a caregiver and their general daily routine. The goal is to get the person to talk about themselves as much as possible. This is why impersonal questions about work were chosen so as not to embarrass the interviewee. This exercise also allows the subject to gain confidence.

2.2.3. Professional care task

The experiment consists in recording speech during the staging of a care task with a fictitious unresponsive patient. The chosen care task is dressing up, which includes buttoning a shirt and the techniques of waking up the body. At the end of the dressing, the caregiver performs an uprighting of the fictitious patient and makes them walk. The fictitious patient is completely mute – which allows to evaluate the technique of *auto-feedback* on the part of the caregiver – and lets themselves be cared for by the caregiver. For the care task, the context was set up to create a familiar environment for the caregiver. The bed used is similar to that of the regular patients and partitions are put in place to create intimacy with the fictitious patient.

2.3. Collected data

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

We recorded a total of 25 participants during the three sessions : 21 female participants and 4 male. Given that we only had 4 male participants, we decided to focus on female participants in this study. One recording of a female participant was excluded due a poor recording quality. The total duration of remaining 20 participants' recordings is 2 hours and 30 minutes. The total duration recorded for each task and the mean duration per speaker per task is 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 phonetic transcription

Phonetic transcriptions were obtained by using an automatic system based on the Kaldi framework [11] and trained using the ESTER database [12]. The model used for the transcriptions is a *Time Delay Neural Network* coupled with a hidden Markov model. The neural network is based on a time-delayed subsampled neural network (TDNN) with 7 TDNN layers and 1024 units in each. The input for the acoustic model is a 40-dimensional high-resolution MFCC vector concatenated with a

100-dimensional I-vector [13].

3.2. Parameter extraction

Automatic segmentation in Inter-Pausal Units (IPUs) is carried out using the output of the automatic transcription system with a threshold of 250 ms. Some shorter pause segments were still present after the automatic segmentation in IPUs. The IPUs were therefore further divided into smaller IPUs between pauses. Hereafter, the IPU refers to the smaller IPUs between pauses.

Fundamental frequency (F0) and intensity were extracted every 10 ms using the snack routines [14]. Mean and standard deviation of the F0 and intensity values are computed on each IPU. The fundamental frequency was converted to semitones (ST) relative to a reference value (here, 50 Hz).

The speech rate was computed on each IPU using the phones detected by the phonetic transcription system and is therefore given in phones per second.

4. RESULTS

Linear mixed models (LMM) were used to analyze the acoustic parameters using the *lme4* package in R [15]. A model is conducted for each acoustic parameter. Speech style was included as a fixed factor for all models. For the models on IPU duration and speech rate, intercepts were included for subject as random effect. Concerning the models on fundamental frequency in semitone, the standard deviation of fundamental frequency across IPUs and the intensity, intercepts were included for subject and item as random effects.

Figure 1 presents the duration of the IPUs (in seconds) as a function of the three speech styles. The results show that IPUs on average are shorter for the interview [β = -2.210; t = -2.87; SE = 0.771] and reading tasks [β = -3.334; t = -4.32; SE = 0.772] than for professional voice.

This observation is in line with what is expected by the use of the *auto-feedback* technique in which the caregivers are asked to keep a steady vocal flow through the length of the care.

This lengthening of IPUs in the professional task is linked to a lower speech rate, as seen on Figure 2. 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 in line with our expectation, 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 a indication of a calmer voice.

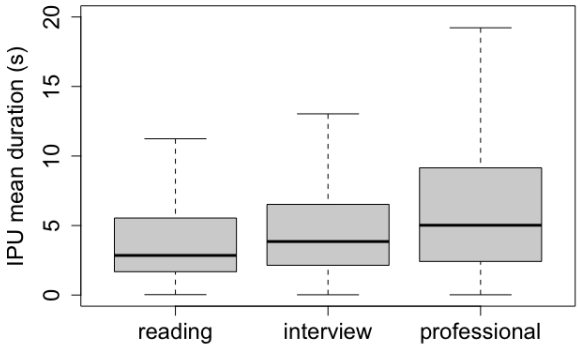

Figure 1: Boxplot of IPU length in seconds for the three speaking styles

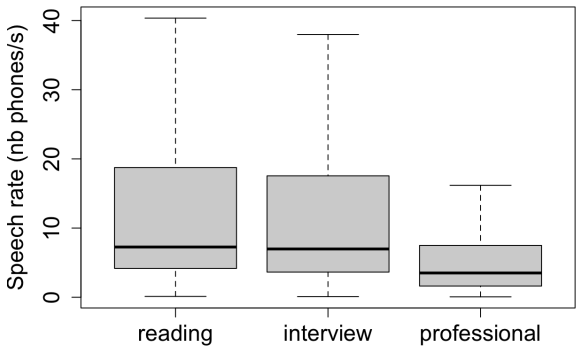

Figure 2: Boxplot of the Speech Rate in phones per second for the three speaking styles

Figure 3 illustrates the fundamental frequency (in semitones) as a function of the three speech styles. We observe that fundamental frequency is on average lower for the informal interview [β = -2.797; t = -2.71; SE = 1.034] than for the professional voice. No significant difference is found between reading and the professional voice. This finding implies that the caregiver try to speak with a significantly higher voice following the instruction of the "Humanitude" care method during a professional setting.

The melodic variations in terms of standard deviation of the fundamental frequency across IPUs are displayed on Figure 4. This figure shows that caregivers produce less variation in fundamental frequency when interviewed than when they are doing professional care [β = -0.210; t = -1.25;

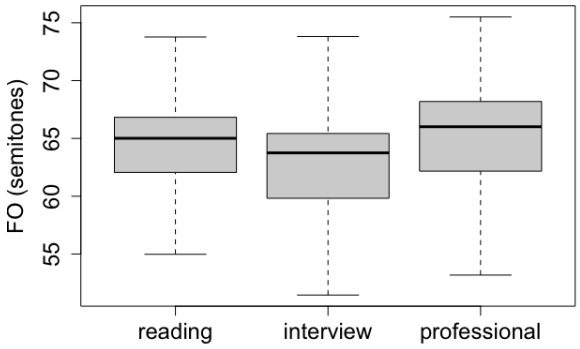

Figure 3: Boxplot of Fundamental Frequency values average over IPUs for the three speaking styles

SE = 0.168]. As far as the standard deviation of the fundamental frequency is concerned, no significant difference is found between reading and the professional voice, based on the relevant LMM model. Having greater variations in their use of fundamental frequency is corresponding to the expectations of producing a melodic voice as recommended by the Humanitude care method.

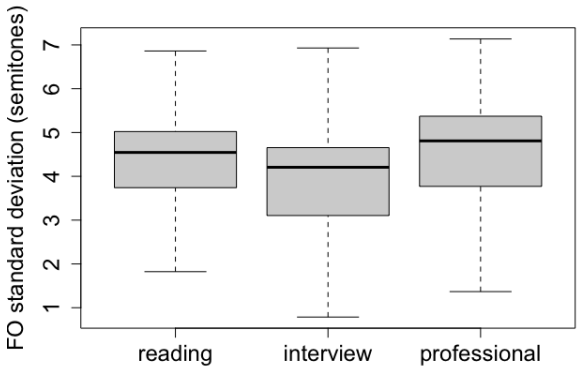

Figure 4: Fundamental frequency standard deviation across IPUs in semitones for the three speaking styles

Figure 5 shows the measured intensity (in dB) as a function of the three speech styles in question. The intensity is observed to be higher for the interview [β = 2.98; t = 2.48; SE = 1.20] and reading [β = 4.78; t = 3.98; SE = 1.20] conditions than for the professional voice. This finding also corresponds to our presumptions since it relates to the fact that caregivers are expected to speak in a calm voice.

The differences observed on intensity for the three different styles may also explain the fact that our speakers tend to produce higher pitch speech when caring because it compensates for the loss in intelligibility induced by the lower intensity.

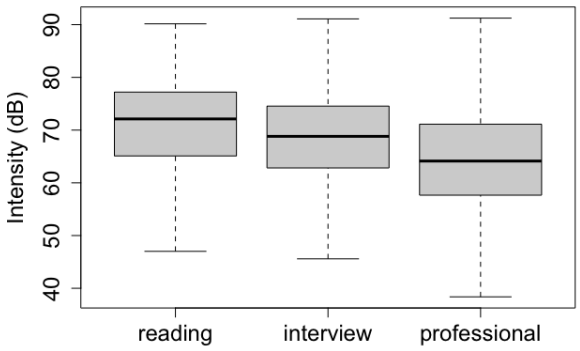

Figure 5: Intensity in decibels for the three speaking styles

5. DISCUSSION

This paper aims to identify the characteristics of caregivers professional speech in contrast with their production while reading or performing an informal interview. As stated in the introduction, caregivers are recommended to produce a soft, calm and melodious voice.

The three prosodic parameters (i.e. fundamental frequency, intensity, duration) we studied here are all in line with this recommendation. For instance, we observed longer IPUs, lower speech rate, greater F0 variation, higher F0 and lower intensity during professional care than during interview or/and reading. The "soft voice" characteristics may be related with this lower intensity and higher F0 voice instructed by Humanitude method. This Â« softness Â» of the voice may also be related to other acoustic parameters like voice quality in interaction with fundamental frequency and intensity. Further studies are needed to confirm or refute this hypothesis.

The next step of our research will focus on the lexical content of the caregivers' speech. It is indeed expected that their professional interactions convey positive emotions. We plan to investigate this assumption by using the transcriptions generated by a high-performance end-to-end automatic speech recognition system coupled with lexical and sentiment analysis tools.

6. ACKNOWLEDGMENTS

This work has been partially funded 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] 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/