Harnessing the Role of Speech Interaction in Smart Environments Towards Improved Adaptability and Health Monitoring

Fábio Barros[1,2], Ana Rita Valente[1,2], António Teixeira[1,2], and Samuel Silva[1,2(✉)]

[1] Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, Aveiro, Portugal
{fabiodaniel,rita.valente,ajst,sss}@ua.pt
[2] Department of Electronics, Telecommunications and Informatics (DETI), University of Aveiro, Aveiro, Portugal

Abstract. The way we communicate with speech goes far beyond the words we use and nonverbal cues play a pivotal role in, e.g., conveying emphasis or expressing emotion. Furthermore, speech can also serve as a biomarker for a range of health conditions, e.g., Alzheimer's. With a strong evolution of speech technologies, in recent years, speech has been increasingly adopted for interaction with machines and environments, e.g., our homes. While strong advances are being made in capturing the different verbal and nonverbal aspects of speech, the resulting features are often made available in standalone applications and/or for very specific scenarios. Given their potential to inform adaptability and support eHealth, it's desirable to increase their consideration as an integral part of interactive ecosystems taking profit of the rising role of speech as an ubiquitous form of interaction. In this regard, our aim is to propose how this integration can be performed in a modular and expandable manner. To this end, this work presents a first reflection on how these different dimensions may be considered in the scope of a smart environment, through a seamless and expandable integration around speech as an interaction modality by proposing a first iteration of an architecture to support this vision and a first implementation to show its feasibility and potential.

Keywords: speech interaction · nonverbal speech features · health monitoring · eHealth · multimodal architectures

1 Introduction

In a world where we are surrounded by technology, researchers have been using assistive technology to build intelligent environments around people aiming to

This work is partially supported by FCT grant 2021.05929.BD and by IEETA - Institute of Electronics and Informatics Engineering of Aveiro Research Unit funding (UIDB/00127/2020).

A. Cunha et al. (Eds.): MobiHealth 2022, LNICST 484, pp. 271–286, 2023.
https://doi.org/10.1007/978-3-031-32029-3_24

increase their quality of life and also provide a protected and secure ecosystem to assist them in their daily living (e.g. Smart Houses) [6,23,28]. This assistance can entail support for a wide variety of tasks, e.g., controlling equipment in the household or improve how interaction can be performed with different technologies to support communication with family and friends, and can also be complemented with the monitoring of the persons health condition informing, for instance, how the person might need to be motivated to perform more exercise or if attention from a health professional is warranted. There are several ways of monitoring the person's physical and cognitive state. However, approaches that are noninvasive and part of the environment, instead of equipment that needs to be worn are desirable, since they are less intrusive and do not depend on user adherence (or caregiver intervention) [34]. In this context, data that can be collected from normal user activity, during the day and night could play an important role. Examples of such data can be posture, mobility, gestures, facial expressions, and also speech.

Speech, our most natural and efficient form of communication, is currently one of the most promising modalities to interact with technologies and environments due to a strong development of speech technologies in recent years (boosted by the availability of services for speech recognition, e.g., from Google and Azure). In this context, most of the works applying speech technologies for interaction with smart environments have been including services that perform speech recognition, speech synthesis, or even dialog systems [33]. And speech can be an important and intuitive form of interaction for all, but its consideration may be particularly important in assistive scenarios, since it potentially improve adaptability to different audiences. Additionally, speech is more than just words and a strongly personal trait, and how it is produced can add to the communication, e.g., intonation, stress and prosody [12], and reflect aspects of our physical and mental condition. In this regard, it can be harnessed to understand the emotional and cognitive state of humans and, in recent years, it has been widely explored for the detection, prevention and follow-up of some diseases, disorders, and communication problems, such as Alzheimer's or bipolar disease, through the analysis of semantic [35] and acoustic features [15].

All this wealth of information that can be obtained from speech has motivated intense research on its different dimensions and the literature has been prolific in providing methods to do so [10,11,22,26]. Nevertheless, the nonverbal aspects of speech along with its value as a biomarker for several conditions, which could play a pivotal role in increasing system adaptiveness and adding to the monitoring capabilities of assistive environments, are still not widely considered. In this regard, solutions that provide these features as part of individual applications, as is often the case, while showcasing the value of these technologies, narrow the scope of their use, and bringing them to interactive environments in a way that fosters their utility to a broader set of services is strongly desired. Furthermore, all these methods entail a certain degree of complexity and, from a developer's perspective, having to master the integration and use of several of these methods may be troublesome. Therefore, it would be important to bring these methods into interactive assistive environments as off-the-shelf features that different

services could just profit from, e.g., to improve adaptation, or understand more about the user's physical and mental status. Additionally, with the increased consideration of speech interaction, the analysis could be performed over the daily interactions with the system, adding to the naturalness and non-intrusiveness of how the monitoring is performed. In this context, the work presented here provides a first proposal of how this integration can be performed reflecting on an initial set of requirements, proposing the overall architectural aspects to serve this purpose, and describing a first proof-of-concept implementation showcasing our proposal.

2 Speech Communication and Interaction

In what follows we provide a short overview of the potential of speech data to improve both interaction and health-related monitoring. This does not aim to be a thorough account of all possibilities, but an illustration of the kind of features we aim to integrate.

2.1 Intrinsic Properties and Applications of Speech

Speech is a form of communication that is, often, closely associated with the spoken words. However, there are also nonverbal cues present in speech that provide more information than just the linguistic content [9,16,25].

By considering them we can, in fact, potentially make human-machine interaction easier and more satisfying, increasing its credibility and realism, and increasing feelings of empathy and affection [5]. Methods such as speaker emotion recognition [1], speaker verification/identification [2], and detection of language, age, and gender [14,20,31] are some examples illustrating how can speech contribute to a more adaptive human-machine interaction. Furthermore, several nonverbal speech cues, i.e., vocal sounds without speaking any word, can express important information for the communication process, for instance by informing we are paying attention or that we do not agree with what is being said.

In addition, the acoustic and linguistic resources of speech are also a valuable resource for healthcare. They become important in the detection, prevention and intervention of neurological and psychiatric disorders such as Autism, bipolar disorder, Schizophrenia, Alzheimer and others [16]. There are several studies indicating its importance, for example, Tanaka et al. [29], conducted a study of speech characteristics, such as pitch, intensity, speech rate, and voice quality for the detection of autism. Karam et al. [19] extracted a set of low-level features using the openSMILE toolkit [11] such as RMS energy, zero-crossing rate, pitch, voice activity detection, mel spectrum for a long-term monitoring of mood states for individuals with bipolar disorder. Sch-net, proposed by Fu, Jia, et al. [13], by extracting fluency, intensity-related and, spectrum-related features, achieved a deep learning architecture, for detection of schizophrenia. Additionally, voice tone alteration, speech rate alteration and speech speed alteration are features characterizing Alzheimer's disease according with Bertini et al. [4].

Overall, these works show a very recent, diverse, and active work on harnessing the richness of speech with a wide variety of approaches. Nevertheless, one relevant aspects is that these methods are often still not deployed in a wider application scope or they are made available through standalone custom applications, e.g., to support a specific clinical setting. Our vision is that if these can be brought to profit from the growing role of speech as an interaction modality, their applicability is widened and their potential is driven farther towards eHealth.

2.2 Supporting Speech Interaction

Considering that we are evolving towards interactive environments with am increasingly dynamic number of devices, applications, and interaction modalities, such as smarthomes, architectures that can support interaction in these transient environments are paramount. However, there is no standardization of these technologies and methods so that existing efforts to support multimodal interaction can be reused in different scenarios beyond those for which they were originally designed. To tackle these adversities, several efforts have been made supported on the recommendations of the W3C for multimodal interactive architectures [8], such as Mudra [17], and Cu-me [30]. Among these efforts, a representative example is the AM4I architecture and framework [3]. It is based on a modular design where interaction modalities (e.g., speech, gestures, touch) are decoupled from applications. Besides this decoupling, the provided framework also has the advantage of providing some modalities off-the-shelf. This is important since it means that a developer, when integrating a novel application, does not need to master any of the interaction technologies and if different modalities or devices are available, at different times, this is transparent to the applications. A notable example of how this can be an advantage is the availability of what the authors call a generic speech modality, i.e., a modality that already tackles all aspects regarding speech interaction, e.g., speech recognition, and grammar definitions, making it simpler to use.

In light of how these multimodal architectures allow dealing with the modern interactive environments, such as smart homes, we consider that they should work as a reference context for our proposals.

2.3 Discussion

There are a number of important advantages that can come from enabling a stronger role of speech data beyond speech recognition, in smart environments, particularly when the integration of adaptability and health state monitoring are indisputable goals, such as in those scenarios aiming to provide assistance to daily life tasks while keeping an eye on the physical and mental health:

– advancing our knowledge and support to nonverbal cues in speech interaction is a paramount aspect to consider in our communication with interactive systems as it can potentially improve efficiency, naturalness, and adaptiveness;

- the integration of the inherent complexity of tackling nonverbal speech cues as part of the interaction framework, i.e., already residing in the core interaction features supported by the environment, and not as part of what each developer needs to master is paramount to ensure wider adoption;
- if the assessment of the health-related features can be integrated to profit from speech produced during the normal interactions with the home it adds to the naturalness of this process and lessens the feeling of being monitored with potential advantages to the ecological value of what is measured;
- building a long term record of speech features, i.e., measures systematically taken over time, can be important for the detection of changes that might work as biomarkers for early detection of certain physical and mental conditions [7,21,27,32];
- A decoupling among the speech signal, the extracted features, and how they can be used by third party applications opens space for a more versatile management and control regarding privacy of these data. This means, for instance, that a new application added to the smart home environment may profit from data extracted from speech, but without having access to the speech data.

All these aspects considered motivate our vision and proposal as described in the following sections.

3 Towards Enhancing Speech's Role in Smart Environments

Our overall vision is that the speech interactions with a smart environment, such as our home, can be harnessed to provide data for a wide range of services to improve both the role of speech in interaction and its value as a source for health monitoring. To ground our work, we designed a few scenarios to illustrate the overall envisioned features in the context of a smart home specifically designed to provide its users with services to assist in keeping or recovering their health, aligned with ongoing work with industry partners. This brings forward the relevance of adaptive interaction, but also the importance of considering speech for monitoring. While the overall monitoring context for this smart home also entails other modalities, such as, posture and physical activity, we keep to those scenarios harnessing speech, relevant for the work presented here.

3.1 Illustrative Scenarios

The provided scenarios illustrate two different audiences that can profit from the envisaged features: the user, at home, to obtain a more adaptive response from the system, and a caregiver or health professional, to be able to monitor for particular conditions, over time, or receive a notification about a notable event.

In a first scenario, Rosa is at home and the system is able to infer some instability and take some action to understand the situation, help Rosa handle it, and monitor the outcomes.

Scenario 1: Rosa having a different day—Rosa is a 76-year-old female who is a health smart home patient. Rosa is a well disposed, calm and active person, but she has heart and anxiety problems. On a particular day, the health smart home realizes that something unusual is going on with Rosa, through her continuous monitoring and comparison of her user context model, unusual values have been detected in her interactions since her speech rate as lower then the normal and as well as the emotion associated are classified as "sad". The conversational agent at the health smart home proceeds to interact in order to ask Rosa and understand what might be the reason for the problem. She responds by informing it that she is feeling a little anxious. The assistant then advises some diaphragmatic breathing and guides Rosa in how to proceed. After this, the system asks Rosa about how she his feeling now and detects less distress and a more positive attitude.

In a second scenario, Rita, who is Rosa's caregiver, is notified of some changes in Rosa's behavior and tries to understand what is happening, in a first instance, analysing some data about the past days.

Scenario 2: Rita follows Rosa's Day Remotely—Rita is a 38-year-old female who is the health smart home caregiver. In a special way, Rita is the caregiver responsible for Rosa, who is a patient at the health smart home. Rita, although a very attentive caregiver, can't dedicate 100% of her time to her patient. On Monday morning, Rita is notified by the system that Rosa during the weekend made use of multiple interactions with a negative sentiment associated with her speech, a high speech rate as well as an emotion associated with "anger" which are completely different values from Rosa's context model of a patient who is a well disposed, calm and active person. Rita checks that the assistant already tried to suggest some breathing exercises without any long-lasting impact. In this sense, and in order to evaluate and analyze these interactions herself, Rita searches for all the interactions made by Rosa during that weekend. The result of this query is returned to her on a dashboard containing different parameters about the queried period, including voice parameters, semantic graphs, and word clouds of most used vocabulary so that she can evaluate when and how Rosa's situation changed and what dimensions are affected. She notices that words such as "alone" and "pain", came up more frequently than usual, in the last couple of days. Rita goes out to visit Rosa.

One important aspect to note is that the overall idea is not to provide the third-parties with a complete account of the dialogues the person had, at home, but a set of aggregate data. And different people will only have access to data with relevance for their role. Access to complete sentences, for instance, may be possible, but solely for very well defined contexts, e.g., a remote consultation with a therapist. This overall principle is intended as a first rule to preserve some degree of privacy. Additionally, the system is not intended to be the last instance of diagnosis, but rather a tool that allows and facilitates the collection

of additional information, in other words, the system serves only as a first line of indication for the therapist not a diagnostic tool.

3.2 Overall Requirements

In light of the proposed vision and scenarios, there are a few notable high level features that need to be considered:

- adopt a modular design, given the diversity of methods that can be relevant for integration, also enabling expansion, over time, to serve new purposes and integrate novel methods;
- obtain data from different sources of speech audio (and, eventually, written speech) in the environment;
- integrate with multimodal interaction architectures that have already been adopted to support the interaction in increasingly complex interactive environments entailing multiple services, devices, and interaction modalities;
- make the outcomes of the different methods available to a broad set of services that might not know how to deal with specific results and need a higher level (semantic) outcome;
- support different timespans for analyses: (a) as a long-term repository, to provide a baseline and as an history of parameter evolution; (b) as a source for overall characterization over smaller periods of time, e.g., daily mood or status; and (c) as a source for immediate system reaction, e.g., adapting to a particular emotional change;

These overall requirements inform our first proposal for a broader consideration of speech-related features serving adaptivity and monitoring as described in what follows.

4 Development

We start by the proposal of the overall architecture that should address the identified overall requirements. At this first stage, and as further explained ahead, we did not go into full detail regarding a complete integration with multimodal interactive frameworks, since this is not essential to demonstrate the feasibility of our proposal. After, we detail the first instantiation of this architecture and showcase some of its potential.

4.1 Proposed Architecture

Figure 1 shows the overall architecture of the proposed system, as explained in what follows.

Core Modules. The audio stream can come from a wide variety of microphones integrated in the house and can be accessed by the Speech Processing module. This module is responsible for extracting both verbal and nonverbal (e.g., words, speech rate, fundamental frequency) data. The purpose of this module is solely the computation of different features from the speech data and their storage, and no conclusions about what the data might mean or if it requires some action are drawn, at this point.

To make sense of the data extracted by the processing module, the Analyzer module provides features to further process and interpret the data according to particular goals. This can be done by resorting to the data provided by the processing module, for different time windows (current or past), along with data from the user context providing reference values or states. This module may enable, for instance, detecting a change in a voice property of interest or updating the user context with relevant information, such as the current mood. The analyser module can even implement more complex logic to, for instance, perform semantic analysis based on transcriptions of the speech data, with known relevance, e.g., through semantic graphs analysis, to monitor cognitive aspects [24].

Finally, the system integrates the Information and Insights module that supports providing relevant information from what has been collected/analyzed, e.g., a dashboard exhibiting interactive visualizations of data required by the caregiver/healthcare professional to assess the patient's state. This module can provide a set of off-the-shelf solutions for visualization and features allowing the inclusion of new ones. It profits from the Data Inspector module for querying the available data.

Integration with Multimodal Interactive Framework. The architecture in Fig. 1 also depicts, delimited by a dashed line, some elements that should already be a part of the existing interaction infrastructure, in the environment, including applications and those components belonging to the multimodal interactive framework, e.g., the interaction manager and interaction modalities [3], with just speech modalities represented, for the sake of simplicity.

The Assistant element represents an application, e.g., a home assistant, working in the interactive environment. It is this application that receives the interactions from the users through speech (and other modalities) and can serve a wide range of purposes, such as enabling them to control equipment in the house or ask questions about its functioning.

In a first instance, our proposal can be integrated with existing multimodal interactive frameworks, such as AM4I [3], through the user context, providing the interaction modalities and applications (such as the Assistant) with data potentially informing system behaviour and adaptability. Naturally, deeper levels of integration may be devised, over time, e.g., with analysers being able to request the assistant to initiate an interaction to obtain more speech data from the user or clarify context, e.g., to confirm if the user as a cold (which could affect the voice).

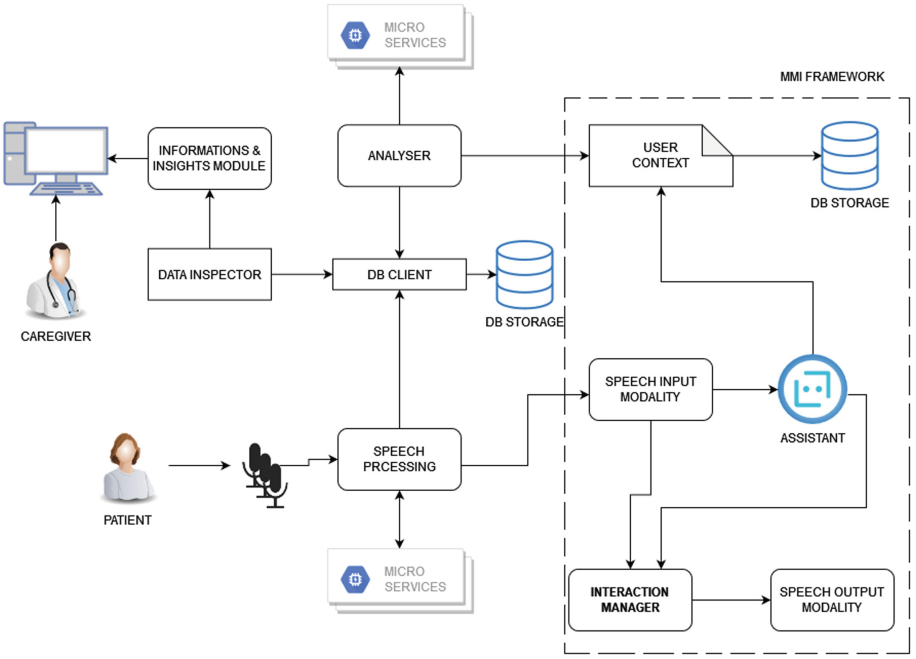

Fig. 1. Overall architecture to address the identified requirements- The modules to the right, inside the dashed area, represent the overall components of a multimodal interaction framework to illustrate the points of contact/integration with the current stage of our proposal.

It is also important to note that the Speech Input Modality already present, in the environment, as part of the framework supporting multimodal interaction, can also be an important point for integration. The diagram in Fig. 1 shows two alternatives. A first alternative is that the speech modality obtains the audio stream as before, bypassing the speech processing and analysis modules, and does not need any change. If it already had some logic to deal with context changes, it can profit from them, as explained above. This is the simplest level of integration, requiring no changes to the framework. The second alternative is that the speech input modality receives speech data that has already gone through some processing/analysis and thus provides more metadata for the speech modality, e.g., which word was emphasized. This second option is the desirable long-term integration method as it enables nonverbal speech cues to be more widely explored during interaction.

In the following sections, additional details are provided about the concrete technical aspects regarding a first implementation of the different architectural modules.

4.2 Data Management

One important aspect of the proposed architecture pertains how to deal with the speech data and all the data that is subsequently produced by the different modules. In this regard, there are some aspects that need to be considered:

- there is no speech data continuously pouring into the system, which means that the selected approach needs to be driven by data availability;
- given the modularity of the proposal, multiple methods will need to access the data in parallel;
- for particular analyser methods, only a subset of data made available by processors will be of interest from the whole range that is produced, meaning that a data source selection mechanism should enable this focus;
- some methods may need to provide results on the currently arriving data, while others may do so only at certain times, e.g., over a day, and based on past data.

Aligned with these needs, our approach adopts a producer/consumer approach based on streams using Apache Kafka[1]. This allows for the different modules to subscribe to the topics of interest and an elegant way to manage the asynchronous nature of data availability and production.

DB Storage is a non-relational and document-oriented database that uses MongoDB[2]. The database is responsible for keeping the data extracted from each speech sentence and data related of the user context. As a non-relational and document-oriented database this allow us to store variations in the documents structure and also storing documents that are halfway complete.

DB client is a module responsible to store and retrieve the data extracted and stored providing a versatile way to inspect the wide range of available data according to the needs of each analyser, e.g., some may require more than one parameter, others may work on different time windows.

Finally, the data required to build the required dashboards is supplied by a **Data Inspector** developed in Flask that queries the stored data through the DB client.

4.3 Speech Processors

The Speech Processing module adopts a micro-services based implementation with the purpose to be used according to the needs of the system, but also, because it is possible to include or remove micro-services without affecting how the module integrates with the remaining architecture acting as a hub of micro-services. For this first implementation, we wanted to deploy processing services actuating at different levels of the speech data (acoustic and language) entailing:

- **Speech-To-Text:** This micro-service transforms the audio stream into words and also the segmentation of it, i.e., given a word, extract its duration, and the

[1] https://kafka.apache.org.
[2] https://www.mongodb.com/.

instant when it occurred. To perform this transcription we use the "Speech to Text" service provided by Microsoft Azure.
- **Speech Rate:** This micro-service calculates the number of words spoken per minute.
- **Speech Emotion Recognition:** This micro-service computes the probability associated with different emotions (in a set of 6) from the audio stream.
- **Sentiment Analysis:** This micro-service classifies each spoken sentence into "Positive", "Neutral" or "Negative" with a probabilistic value. To perform this classification we integrate the "Azure Text Analytics" service provided by Microsoft Azure.
- **Audio Features:** This micro-service extracts a set of audio features, such as fundamental frequency, intensity, and energy, from the audio stream integrating the OpenSMILE library.

4.4 Speech Analyzers

The Analyser also adopts a micro-services based implementation and is responsible for retrieving the data stored and analyse it for specific goals, e.g., if pitch changed significantly, or to compute the monthly baseline speech rate to populate the user context with a reference value. Furthermore, this module is also responsible for triggering actions, e.g., alerting the assistant and the caregiver if there are anomalous values in the interaction/monitoring. At this stage, the analyser implements three micro-services, which are:

- **Sentiment Analyser:** This micro-service compares the classification and probability of the spoken sentence with the user overall expressed sentiment in a specific time period, e.g., the past hours or week.
- **Emotion Analyser:** This micro-service monitors for emotion changes given a period of reference or the mood set in the user context.
- **Speech Rate Analyser:** This micro-service checks, using the the user context model or data from a reference time period, if there were very abrupt speech rate variations.
- **Acoustic Features Analyser:** This micro-service, at this stage, and for illustrative purposes, can understand if there were many variations in the fundamental frequency.

All the data considered by the analysers is generated by processing microservices or other analysers and selected according to the analyser's needs. Therefore, the inclusion of a new analyser micro-service may also entail the integration of novel processing features. On the other hand, some analysers may just profit from already existing data.

4.5 Assistant and User Context

The Assistant and User Context depicted in the architecture represent elements belonging to the framework dealing with multimodal interaction with

which our proposal aims to integrate. Since our purpose, at this stage, was the deployment of the core features described in the previous sections, as a proof-of-concept, these two elements have received minimal intervention. The assistant inherits from our ongoing work with smart homes and allows controlling the house's lights and appliances, and provides information about water and energy consumption. Its sole purpose for the scope of this work was to provide a motive to interact with the house and, thus, generate speech data.

The User Context is an important element, in our proposal, since it is a first path to foster integration of the work presented here with the multimodal interactive framework. By populating the user context with relevant information about the user, e.g., current mood, emotional changes, or speech rate, – through the analyser module features – the interactive services, e.g., the assistant, may be able to adapt their functioning or behavior accordingly. This integration method keeps the interactive services agnostic to all the processing and analysis that is being performed in our proposal.

4.6 Visualization Dashboard

The Information & Insights module supports providing third parties such as caregivers, health professionals or, researchers with ways of interactively exploring relevant data to inform assessing the user status, e.g., looking for any abnormal change. To this end, we required an approach that could easily allow the visualization of different types of data, over time or even, consume the data directly from streams, e.g., if a more immediate visualization is required, e.g., during remote consultation with a Speech and Language Therapist. For this purpose, different solutions were considered, such as Kibana[3], Grafana[4] and the development of dashboards from scratch. While all provided the overall features required, a dashboard developed from scratch that provides much more adapted visualization and versatile, was the selected approach.

To demonstrate the envisaged functionality and usefulness, we developed a minimal dashboard that illustrates how a healthcare professional/language pathologist might have access to relevant data concerning the patient's monitoring.

Figure 2 illustrates a simple dashboard built using ReactJS[5] showing different information concerning the suprasegmental characteristics of the speech resulting from reading a small text collected from the fable "The North Wind and the Sun" which is phonetically balanced [18]. It includes data on parameters such as the speech rate, the fundamental frequency contour (intonation), or the intensity trace. For a Speech and Language Pathologist, this type of information can be used to, e.g., characterize the speech of an individual through the comparison with normative/baseline data, or to assess the effectiveness of an intervention process. For instance, if an abrupt and sustained decrease is observed in the

[3] https://www.elastic.co/pt/kibana/.
[4] https://grafana.com/.
[5] https://reactjs.org/.

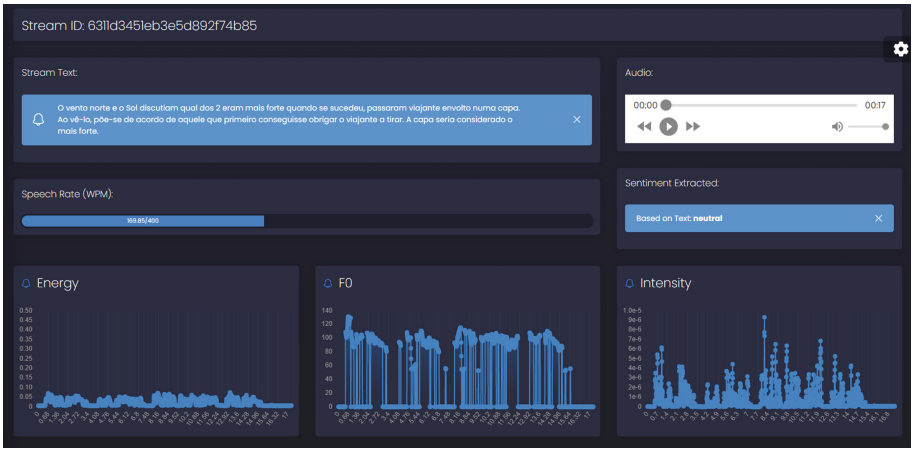

Fig. 2. Dashboard showing different information concerning the suprasegmental characteristics collected from a user reading the fable "The North Wind and the Sun" such as: speech rate, energy, fundamental frequency and, intensity.

fundamental frequency plot, it may mean the emergence of laryngeal pathology. In addition, if a marked decrease in speech rate is verified, it can also be linked to different communication pathologies or even mood disorders, for example.

While these parameters can be collected from routine interactions with the system, asking the user to read a small text, e.g., to learn it to tell a grandchild – a challenge that can be presented by the assistant –, is also an important aspect that can be explored. In this regard, the literature may provide a strong background to support the speech and language pathologist's assessment from that specific set of data and the person will not need to leave home to have a first checkup or follow-up.

5 Conclusions and Future Work

In light of the importance of speech in our daily life and its growth as a form of interaction with machines, the work presented here argues that the different dimensions provided by speech may play a greater role to inform adaptivity and eHealth approaches. In this regard, we reflect on the overall requirements to address this vision and propose a first iteration of an architecture that should serve this purpose along with a first instantiation, as a proof-of-concept.

Currently, this first iteration of our proposal is integrated in a smart home lab designed to test solutions for monitoring and supporting users with their daily living and health, at different levels (e.g., motor activity, posture, and our proposal for speech). While already enabling the collection of baseline data, and informing on mood changes, this scenario will provide us with a richer context for testing integration and further advance system features, particularly regarding,

at this point, the consideration of nonverbal features extracted from the acoustic speech signal.

In the work presented here, we have implemented the different modules independently from the interaction architecture as a proof-of-concept of our proposal. Nevertheless, and as we argued, these features should progressively become an integral part of a multimodal interactive framework, e.g., The AM4I [3], to maximize their potential and the next stage of the work will progressively address that aspect. In this regard, a stronger integration between the speech input modality and the speech processing module is a step to follow. This will enable a greater integration of nonverbal cues as part of interaction in a more immediate fashion.

References

1. Abbaschian, B.J., Sierra-Sosa, D., Elmaghraby, A.: Deep learning techniques for speech emotion recognition, from databases to models. Sensors **21**(4), 1249 (2021)
2. Abdullah, H., Warren, K., Bindschaedler, V., Papernot, N., Traynor, P.: SoK: the faults in our ASRs: an overview of attacks against automatic speech recognition and speaker identification systems. In: 2021 IEEE Symposium on Security and Privacy (SP), pp. 730–747. IEEE (2021)
3. Almeida, N., Teixeira, A., Silva, S., Ketsmur, M.: The AM4I architecture and framework for multimodal interaction and its application to smart environments. Sensors **19**(11), 2587 (2019)
4. Bertini, F., Allevi, D., Lutero, G., Calzà, L., Montesi, D.: An automatic Alzheimer's disease classifier based on spontaneous spoken English. Comput. Speech Lang. **72**, 101298 (2022)
5. Bozkurt, E., Yemez, Y., Erzin, E.: Affective synthesis and animation of arm gestures from speech prosody. Speech Commun. **119**, 1–11 (2020)
6. Calvaresi, D., Cesarini, D., Sernani, P., Marinoni, M., Dragoni, A.F., Sturm, A.: Exploring the ambient assisted living domain: a systematic review. J. Ambient. Intell. Humaniz. Comput. **8**(2), 239–257 (2017)
7. Chojnowska, S., Ptaszyńska-Sarosiek, I., Kępka, A., Knaś, M., Waszkiewicz, N.: Salivary biomarkers of stress, anxiety and depression. J. Clin. Med. **10**(3), 517 (2021)
8. Dahl, D.A.: The W3C multimodal architecture and interfaces standard. J. Multimodal User Interfaces **7**(3), 171–182 (2013)
9. Dunbar, R., Robledo, J.P., Tamarit, I., Cross, I., Smith, E.: Nonverbal auditory cues allow relationship quality to be inferred during conversations. J. Nonverbal Behav. **46**(1), 1–18 (2022)
10. Eyben, F., Wöllmer, M., Schuller, B.: OpenEAR-introducing the Munich open-source emotion and affect recognition toolkit. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–6. IEEE (2009)
11. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462 (2010)
12. Farrús, M., Codina-Filbà, J., Escudero, J.: Acoustic and prosodic information for home monitoring of bipolar disorder. Health Inform. J. **27**(1), 1460458220972755 (2021)

13. Fu, J., et al.: Sch-net: a deep learning architecture for automatic detection of schizophrenia. Biomed. Eng. Online **20**(1), 1–21 (2021)
14. Garain, A., Singh, P.K., Sarkar, R.: FuzzyGCP: a deep learning architecture for automatic spoken language identification from speech signals. Expert Syst. Appl. **168**, 114416 (2021)
15. Guidi, A., et al.: Voice quality in patients suffering from bipolar disease. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6106–6109. IEEE (2015)
16. Hampsey, E., et al.: Protocol for rhapsody: a longitudinal observational study examining the feasibility of speech phenotyping for remote assessment of neurodegenerative and psychiatric disorders. BMJ Open **12**(6), e061193 (2022)
17. Hoste, L., Dumas, B., Signer, B.: Mudra: a unified multimodal interaction framework. In: Proceedings of the 13th International Conference on Multimodal Interfaces, pp. 97–104 (2011)
18. Jesus, L.M., Valente, A.R.S., Hall, A.: Is the Portuguese version of the passage 'The North Wind and the Sun' phonetically balanced? J. Int. Phon. Assoc. **45**(1), 1–11 (2015)
19. Karam, Z.N., et al.: Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4858–4862. IEEE (2014)
20. Kwasny, D., Hemmerling, D.: Gender and age estimation methods based on speech using deep neural networks. Sensors **21**(14), 4785 (2021)
21. Laguarta, J., Subirana, B.: Longitudinal speech biomarkers for automated Alzheimer's detection. Front. Comput. Sci. **3**, 624694 (2021)
22. Lenain, R., Weston, J., Shivkumar, A., Fristed, E.: Surfboard: audio feature extraction for modern machine learning. arXiv preprint arXiv:2005.08848 (2020)
23. Marques, G.: Ambient assisted living and internet of things. In: Harnessing the Internet of Everything (IoE) for Accelerated Innovation Opportunities, pp. 100–115 (2019)
24. Mota, N.B., et al.: Speech graphs provide a quantitative measure of thought disorder in psychosis. PLoS ONE **7**(4), e34928 (2012)
25. Ramanarayanan, V., Lammert, A.C., Rowe, H.P., Quatieri, T.F., Green, J.R.: Speech as a biomarker: opportunities, interpretability, and challenges. Perspect. ASHA Spec. Interest Groups **7**(1), 276–283 (2022)
26. Sanden, C., Befus, C.R., Zhang, J.Z.: Camel: a lightweight framework for content-based audio and music analysis. In: Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound, pp. 1–4 (2010)
27. Schwoebel, J.W., et al.: A longitudinal normative dataset and protocol for speech and language biomarker research. medRxiv (2021)
28. Sun, H., De Florio, V., Gui, N., Blondia, C.: Promises and challenges of ambient assisted living systems. In: 2009 Sixth International Conference on Information Technology: New Generations, pp. 1201–1207. IEEE (2009)
29. Tanaka, H., Sakti, S., Neubig, G., Toda, T., Nakamura, S.: Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 88–96 (2014)
30. Tumuluri, R., Kharidi, N.: Developing portable context-aware multimodal applications for connected devices using the W3C multimodal architecture. In: Dahl, D.A. (ed.) Multimodal Interaction with W3C Standards, pp. 173–211. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-42816-1_9

31. Tursunov, A., Choeh, J.Y., Kwon, S.: Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. Sensors **21**(17), 5892 (2021)
32. Usman, M., Gunjan, V.K., Wajid, M., Zubair, M., et al.: Speech as a biomarker for Covid-19 detection using machine learning. Comput. Intell. Neurosci. **2022** (2022)
33. Vacher, M., et al.: Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation. ACM Trans. Accessible Comput. (TACCESS) **7**(2), 1–36 (2015)
34. Vacher, M., Fleury, A., Portet, F., Serignat, J.F., Noury, N.: Complete sound and speech recognition system for health smart homes: application to the recognition of activities of daily living (2010)
35. Weiner, L., Doignon-Camus, N., Bertschy, G., Giersch, A.: Thought and language disturbance in bipolar disorder quantified via process-oriented verbal fluency measures. Sci. Rep. **9**(1), 1–10 (2019)