SOURCE™

Software Outreach and Redefinition to Collect E-data Through MOTUS

Speech Recognition

Methodological and evaluation report Eurostat
Grant number: 847218 — BE-2018-TUS
Belgium, April 2020

Contact information

Project coordinator – Statistics Belgium

Kelly Sabbe

Statbel, the Belgian statistical office

King Albert-II-street 16

1000 Brussels

+32 (0)2 277 66 30

kelly.sabbe@economie.fgov.be

Beneficiary – Destatis (Statistisches Bundesamt) (DE)

Elke Nagel, statistician

Gustav-Stresemann-Ring 11

65189 Wiesbaden

+49 (0)611 75 8572

elke.nagel@destatis.de

Birgit Lenuweit, senior statistician

Gustav-Stresemann-Ring 11

65189 Wiesbaden

+49 (0)611 75 8572

birgit.lenuweit@destatis.de

Subcontractor – MOTUS developer

Joeri Minnen

hbits CV – Spin Off Vrije Universiteit Brussel

Witte Patersstraat 4

1040 Etterbeek

Tel: +32 497 18 95 03

Joeri.Minnen@hbits.io / Joeri.Minnen@vub.be

CONTENTS

List of figures

(hbits.9) Speech recognition as a new mode

Deliverable	An exploration on how (future) tools can support speech recognition to broaden up the method of time-registration.

"Currently respondents type in what they do (Time Use Survey; TUS), or what they consume (Household Budget Survey; HBS). Afterwards this information is post-coded to a database. This is a costly and time consuming procedure. Options today are to work with a detailed pre-defined list of activities or goods (tree select or search select). In the future also speech recognition might be a possibility for respondents to register their activities or expenses. Respondents will then talk to the device and this information will be interpreted via an algorithm to code it into behavior or consumer categories."

Within this report the focus lies on time-use research and needs to be seen as a vision to the future where the arguments are not yet clearly underpinned. In the conclusions this report will come back to the topic of consumption research but within the same innovative idea or framework that is underlying to the time diary approach.

1.1 What is speech recognition?

Speech recognition is the ability of a tool (or engine) to recognize and translate spoken language into text. The recognition itself is based on so-called waveforms, or steady periodic sounds. Subsequently these waveforms are split up in utterances by silences. An utterance is the smallest unit of speech.

These units or utterances are then tried to be recognized by matching combinations of words. However, utterances have several characteristics and depending on different contexts it is highly complex to find an optimal matching solution or model that describes the reality. An optimal solution would take extremely long. Therefore, optimization strategies are applied to find the best probability.

These optimization for speech recognition happens according the following models:

- An acoustic model
- A phonetic dictionary
- A language model

The phonetic dictionary is not very effective. The acoustic model is based on distinct short sound detectors. A difference exists between context-independent models and context-dependent ones. Words are understood to exist out of 'phones'. But these phones have a context. A phone is different depending on the speaker, the style of speech, the device, and so on: the coarticulation can be very different. Part of the models take the context into account, others don't.

A language model calculates the likelihood of a sequence of words and is used to restrict word search. To support the sequential process of matching some words are stripped. Mainly two language models are used:

- N-gram – which use statistics of word sequences
- Finite state – which define speech sequences by finite state automation and weights

The better a model is able to predict the next word (so in the sequence of words), the better the recognition. The more restricted the vocabulary of a model is, the lower the chances.

Mostly the three models (acoustic, phonetic and language) are combined in a tool or an engine to optimize the results of recognition. Tools are oriented to a particular purpose and have a database and vocabulary to where the used (combination of) models are tested.

So, what is needed is (1) input (depending on requirements), and (2) a(n) (adopting) tool that has a defined model and vocabulary (or trained database). These ingredients are being discussed below.

1.2 Basic requirements to include speech recognition in time-use research

Below the basic aspects are listed that are seen as a prerequisite to receive input for speech recognition in the context of time-use research.

These aspects are:

- The data collection devices
- The participation of the respondent
- The quality input it brings to the table

1.2.1 Data collection devices

With the modernization trajectory in mind, a shift is more than likely to keep the time diary online. Several devices can be used for this purpose. The most common used devices are personal computers, laptops, tablets and smartphones. Smartphones are easy to carry and also suitable to keep a diary at regular intervals, while a PC usually has a fixed place and is not taken along during the day. Laptops and tablets (or phablets), depending on their seize, tend to be used as a PC or a smartphone and as such can be seen as an intermediate between both.

While it is quite easy to type text using the keyboard of a PC, keyboards on a smartphone or even a tablet are much more restricted. Even with predefined menu tabs, it is not always easy for respondents to actually fill in their diaries, certainly not when moving from one to the other place. So the main advantage of a smartphone (its constant availability) goes together with one of its main disadvantages: more difficulties to put data in.

Both on a PC and a smartphone it is quite easy to fill in activities and context variables by means of a predefined list, but on a PC it is much easier to type in text if the input has to be done in own words. For both devices it is possible to program tags that lead to suggestions in the predefined list of activities, and since tags usually refer to short words or even a part of a word, typing tags on a smartphone is for most respondents not an issue.

Smartphones are better adapted to the more advanced means of input. A more explicit example is the inclusion of information that is derived from sensors such as an external GPS and wearable sensors and of course also smartphone applications itself. All this is not available on a PC, or not useful if the device is not carried all the time with the respondent.

All devices mentioned have the ability to record the voice of the respondent. But certainly, for Smartphones and small devices speech recognition can be an additional added value to the otherwise necessity to type text on a small screen.

1.2.2 Self-completion on the go

The history of time-use research in Europe shows a large majority for a time diary collection method where respondents fill in a diary themselves. Since the first HETUS-guidelines respondents are asked to complete a Paper-and-Pencil diary for 2 days. One weekday and one weekend day.

Generally it is supposed that the transition from paper-and-pencil to online data collection has a benefit for the quality of the registration. At one hand the quality of the registration is linked to the usability of the tool in registering activities. The easier it is to register an activity, the higher the quality of the registration will be. The quality is on the other hand also linked to the length of the recall period. This length is expressed by the time between doing the activity and registering the activity. The more the registration is done on a regular basis, and as close as possible to the end of an activity, the more limited

the memory decay and so the higher the registration quality. It leads to more exact registrations, in the determination of the beginning and ending time but also in the reporting of small activities that are otherwise more easily forgotten by the respondent.

To support the regular registration of activities, the functionality of speech recognition could be helpful so that respondents can dictate their activities on to go. One of the main advantages of a smartphone is that most people carry it with them all the time. As such, the smartphone is always available to register activities. When this functionality would be effective also the registration period in time use research can be extended to a week or even longer, in comparison to the usual 2 days. This in contrast to the diary method used in the US and the ATUS data collection where a respondent is being asked via the telephone to recall the previous day (called a Yesterday-interview).

1.2.3 Efficient quality checks for fast processing

Another reason to shift between paper-and-pencil to online diaries is the cost of the data collection. Delivering and collecting the paper-and-pencil diaries (by interviewer or by post) is quite complex, time consuming and expensive. In a paper diary, consistency or validity checks are not possible and the coding afterwards is often very time consuming and expensive.

Computer-assisted tools to register the time-use have the advantage that all or most of the input is coded and available in a digital form. Because of this, validity checks can be built in to reduce errors and inconsistencies. As a result, data is collected and transferred much faster, and it is in essence a continuous process. Checks can either be built into the tool (e.g. no registrations possible in the future) or can be triggered by the server (e.g. calculation of the unregistered time).

For speech recognition quality checks and fast processing are a necessity to be able to interact with the device of the respondent in a way that the spoken information is text coded and linked to a predefined activity category.

1.3 The boundaries of speech and voice recognition technologies

Voice recognition technology is believed to become an intrinsic part of most of the future applications. However, it is the accuracy that will decide if this technology becomes an inevitable part in how people will give input to a device.

One of the biggest problems is the immense large variation in how people pronounce wordings. The available software for speech recognition is currently not adequate enough to clearly understand entries, especially if the speaker does not speak a standard language or even a dialect. Another problem is the hardware. To become effective the device being used needs a microphone that is intelligent enough to filter the background noise without losing the natural wordings of the user itself. When speech recognition fails it is most of the time because of the background noise and so out of the user's control (e.g. a noisy train).

Looking to the future speech recognition becomes better when the input is given from new phones and VoIP communication technologies, rather as a mean for mass text input. What is off course interesting is the natural fit of this technology with IoT-connected devices. Voice is meant to become the next frontier of a device interface.

On the internet there are various lists and rankings on voice recognition tools. These lists include:

- Siri – Apple
- Google Assistant – Google
- Alexa – Amazon
- Cortana - Microsoft

Siri is the virtual assistant build in Apple iPhones, Android smartphones have Google Assistant, and Windows uses Cortana. Siri can only be used on an Apple iPhone and Microsoft removed Cortana from Android and iOS on January 31, 2020. Google Assistant can still be used on all smartphones. Furthermore, there are many other alternatives, such as Amazon Alexa, Lyra Virtual Assistant, Dragon Mobile Assistant, Robin, Smart Voice Assistant, Bixby, and many more.

By posting various sorts of questions and commands these technologies are tested and ranked, showing Google Assistant at the top, with Siri and Alexa coming in second and third but with some distance from Google Assistant. But it depends on the internet source and the setup of the test. Although the market is still in full evolution, it is clear that differences in quality among the biggest players in the market – Google Assistant, Cortana, Siri and Alexa - are not that great and rankings change over time.

What, of course, differentiates them is the use of different algorithms, but all of them look for key words and context clues to figure out which commands you wish to activate.

That this market is becoming extremely important and the linkage with IoT is natural is being shown by the acquisition of Snips.ai by Sonos at the end of 2019. Sonos develops and manufactures smart speakers, mostly in the Music industry but with the take-over from Snips also in indoor/interior-automatization. Earlier Sonos was associated with the Alexa and Google Assistant. Snips is experienced in the structured coding of words to operate actions within a defined context (e.g. Air-cooler, lower the temperature in-home to 18°C).

And herein might lie an important difference in the use within time-use research. The purpose of using speech recognition is not to activate something, it is in the first place to text code own wordings of the respondent. Although a combination of both elements is interesting in the sense of

- Activate: 'Open primary activity'
- Voice transcription: 'Go for a run'

the voice transcription can also be started by pushing an icon in the type in box and then start talking as e.g. iOS has this included for his text messages.

All-in-all, with a range of words, used as tags, the setup of a diary with a predefined activity list and a responsive back-office, the app could give suggestions on the screen after the respondent told the app which activity has been done. The respondent then can tap the right activity when it appears. But also here respondents has to do at least three manual operations:

- Give input
- Verify, or Change
- Accept

Working with an output list of multiple activities decreases the error probability. The chance of listing the right activity in a list of 10 selection options is much higher than in a list of 3.

But time use research is more than only the registration of one activity. The data collected within a HETUS time diary holds the following components (see also HBITS.4):

- Begin and end time
- Primary activity
- Secondary activity
- Place
- Transport mode
- With whom
- Computer/device use
- Satisfaction

What is true for a primary activity is also true for a secondary activity. The starting and ending time could just be dictated as well and confirmed by tapping on the screen. The same could work for the context variables: place, with whom, etc.

The problem starts:

- When people refer to multiple activities within one voice recording
- When more tags are picked up and the list of possibilities gets larger
- When the respondent dictates multiple activities
- When a voice recording describes changing contexts
- Or when no information is given about a component at all

A change in one of the components is in time use research seen as a different episode. And episodes are the underlying structures (i.e. rows) of a time diary database. With a 10-minute grain (as is now for the HETUS-diaries there are 144 possible episodes per day). But with a more advanced input the grain determination is hard to define and the minimum grain would then be at the minimum 1 minute resulting in at maximum 1.440 episodes per day.

The answering of the components have a certain flow. To keep the structure of an episode, the app could be programmed so that questions are asked, just like an interviewer would do: what did (are) you do(ing), what was (is) the starting and ending time, did you do something else at the same time, etc.

But at the same time this flow, and the numerous manual actions, makes it difficult to embrace the added value of speech technology within time use research.

The last part of this report before the conclusions provides a suggestion to overcome this.

1.4 Tentative & committed data

From the previous parts of this report it is clear that speech recognition technology is still in an evolving stage. It is getting better, and words (as tags) are still better recognized, and the sequence of the words are being used to frame the context. Summarized, speech to text translation is not yet 100% correct. Reasons are pronunciations of words, languages/dialects, the accepting device but also background noise.

Today the technology is mostly used to control a device based on commands. The playground for time use research is different and more complex. To define an episode also information on the temporal, spatial and social dimensions of an activity is needed. Text coding the verbal description of all the tasks a respondent undertakes will be quite error prone especially when this information is not structured enough to be placed adequately in the tables of a relational database.

Despite the numerous reasons why speech recognition is not that easy to be matched with the philosophy of time data collection, the coding of voice into text and furthermore the processing into pre-coded activities holds some added value as well. At least in part.

Therefore this report proposes a midway solution that is supportive to include (all kinds of different) future developments. The suggestion is to introduce a new solution that supports the registration task of the respondent. This can be done via the plug-and-play method, that is also supportive to the ESS criteria of shareability, comparability, scalability and reliability and has the best chance to be adopted by a candidate platform (see also HBITS.7). In order to achieve this, it is believed to introduce the speech recognition technology as a Microservice. Microservices are small and independent services where every service handles a particular problem or performs a certain task.

A Microservice would run in a separate environment for receiving, processing and verifying the voice recorded data to text codes. It also includes the structuring of the database which itself is linked via a dedicated API to the database of the core environment. The idea behind a Microservice is shown in Figure 1. The way it can be include to a core environment like MOTUS is shown in Figure 2.

Figure 1: Microservice Architecture and Speech recognition as an example of a plugin available via the ESS-shareable platform

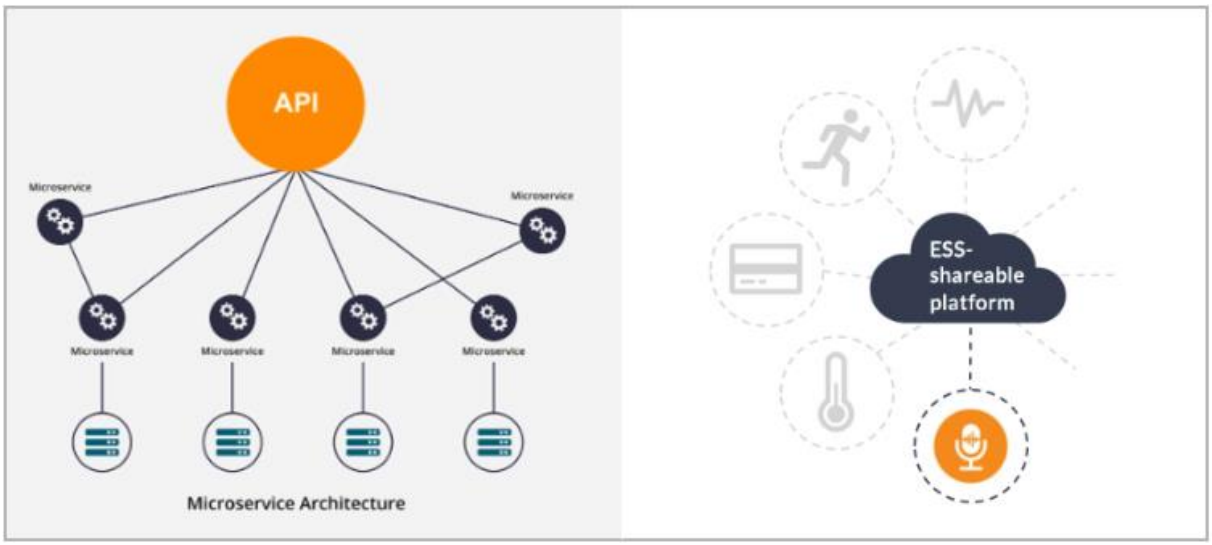
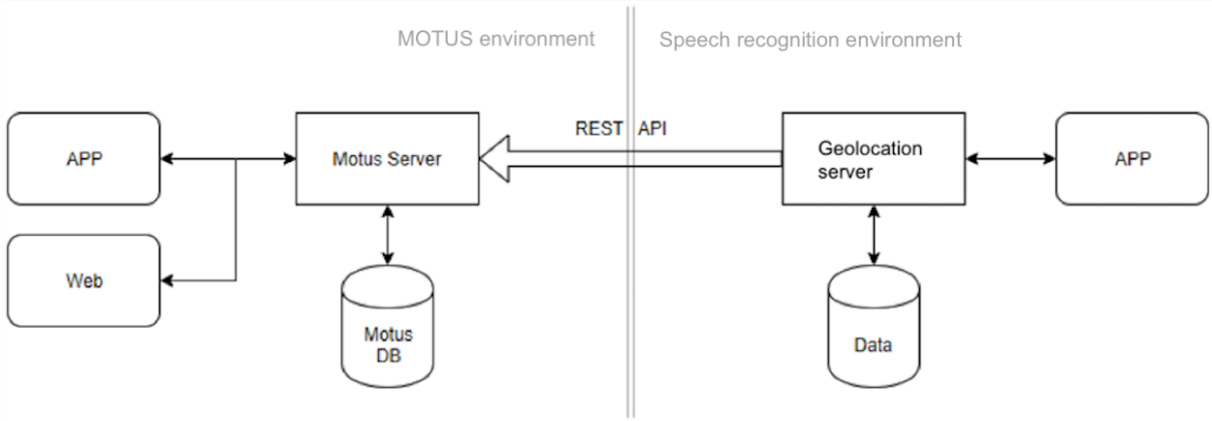


Figure 2: A simplified architectural overview of the relation between a speech recognition environment and the MOTUS CORE environment



The main asset of the Microservice strategy is that:

- It does not influence the operations of the CORE environment
- The input does not have to be error free
- The input can undergo various quality controls
- The output can be parsed
- It is flexible to change from technology, or updates within a technology

Since there is a clear linkage between the input-output and the respondent, the information can be visualized to the respondent on the screen of his device. At that moment the respondent can verify/accept/change the information or extra actions can be run by MOTUS asking for more information (e.g. on the context of the activity). Only at the moment the respondent accepts the information the tentative information becomes committed data.

However, to be effective it needs to be investigated whether or not it is possible to include built-in virtual assistants of the multinational companies into a developed application, e.g. in the same way it is possible to send scanned images to Amazon Textract to extract text and data from an image.

One important aspect to consider in this respect is privacy. It is not clear what the big players like Apple or Google are doing with the information they gather via voice assistants. Open source technologies often give flexible options to developers of apps, but probably have drawbacks in quality. There are many open source project on speech recognition (to give an idea, see https://awesomeopensource.com/projects/speech-recognition), and it needs to be investigated whether they fit in the Microservice strategy where the input does not have to be error free.

1.5 Conclusion

Time research collects rich data but the data collection process is burdensome for respondents, especially when respondents are being asked to keep a record of their activities on the go. When moving more to an online data collection, and the use of Smartphones and applications from the iOS and Android platforms are growing, it is not a surprise that new features to improve the registration by the respondent are investigated.

In this report the technology of speech recognition is been looked at. Given the current state of the art of speech recognition, such as Siri or Google Assistant, an error free implementation into time use research is fairly impossible at this time and for the near future.

This conclusion is also based on the specific organization of how and what data is collected in time use research. Nevertheless, this reports also points on the benefits of speech recognition if included as a Microservice and linked to a CORE data collection environment: data coming from speech recognition does not have to error free, and the respondent has full control over the input. This strategy also has a clear open view for further improvements.

Therefore this reports proposes to investigate the possibilities and chances of success of this strategy, as well as to see if open-source solutions can be used to cover the aspect of privacy as well.

This report did not discuss the topic of speech recognition technology for HBS. However, what is true for TUS is equally true for HBS: the voice of the respondent needs to be text coded, words need to be used as tags, and the context of the words should lead to ranked options from which the respondent can choose. Also here the development solution proposes to use the Microservice strategy to have a shareable, comparable and modular strategy for the future and for the ESS. And with the respondent in the center of the data collection process.