Proceedings of the Language Technologies for All (LT4All) , pages 1–4
Paris, UNESCO Headquarters, 5-6 December, 2019.

Speech Technology for Swedish: Current Impact Areas for Applications and Edyson, an Innovative Tool for Accessing Speech Data

David House, Per Fallgren, Jens Edlund
Division of Speech, Music and Hearing, KTH (Royal Institute of Technology)
Lindstedtsvägen 24, 100 44 Stockholm, Sweden
davidh@speech.kth.se, perfall@kth.se, edlund@speech.kth.se

Abstract

This paper presents four of the impact areas for applications in speech technology identified by the National Swedish Language Bank (Språkbanken) and currently being explored by the Speech Section of the Language Bank (Språkbanken Tal). The four areas defined are Cultural Heritage, Inclusion and Accessibility, Health and Aging, and Digitalization. In addition, the paper introduces Edyson, a tool developed at KTH and used for accessing large quantities of speech data. Current use of Edyson in a research project relating to Cultural Heritage and historical speech recordings is presented and discussed.

Keywords: speech annotation, speech browsing, cultural heritage recordings

Résumé

I detta dokument presenteras fyra viktiga områden inom talteknologi som har definierats av Nationella Språkbanken som angelägna och som kartläggs av Språkbanken Tal. De fyra områdena är Kulturarv, Inkludering och Tillgänglighet, Hälsa och Åldrande och Digitalisering. Dessutom introduceras Edyson, ett verktyg utvecklat vid KTH och som används för att få åtkomst till stora mängder taldata. Nuvarande tillämpning av Edyson i ett forskningsprojekt som rör kulturarv och historiska talinspelningar beskrivs och diskuteras.

1. Introduction

The National Swedish Language Bank (Språkbanken) was inaugurated nearly half a century ago, in 1975, at the University of Gothenburg's Department of Swedish. It has since been a nationally and internationally acknowledged research unit with a focus on language resources and language technology. In 2014, Sweden received funding from the Swedish Research Council to join the European language infrastructure Clarin ERIC, which supports language technology in the humanities and the social sciences, and SweClarin was formed. In 2018, the national research infrastructure National Swedish Language Bank (Nationella Språkbanken) was awarded funding from the Swedish Research Council, which also secured the continuation of Swe-Clarin.

In addition to providing for the continued operation of the original Språkbanken (now Språkbanken Text), Nationella Språkbanken adds two new branches: Språkbanken Sam (Eng. "Society") and Språkbanken Tal (Eng. "Speech"). Språkbanken Sam is operated by the Swedish Language Council at the Institute for Language and Folklore (ISOF), which supports research on the languages, dialects and other parts of the intangible cultural heritage in Sweden; and Språkbanken Tal is operated by the Division of Speech, Music and Hearing at KTH, which caters for resources on speech, speech science, and speech technology.

This paper presents four of the impact areas identified as important for the development of applications using speech technology in Swedish: Cultural Heritage, Inclusion and Accessibility, Health and Aging, and Digitalization. Examples of current activities in each area are presented including a more detail description of the tool Edyson, used for accessing large quantities of speech data specifically within the area of Cultural Heritage.

2. Four impact areas

2.1 Cultural Heritage

Cultural Heritage is an important area both on a national and a European scale**:** according to EU Commissioner Carlos Moedas (in charge of Research, Science and Innovation) positions innovation in cultural heritage in "*the intersections between old and new, between physical and digital, and between disciplines*". The European Strategy Forum on Research Infrastructures (ESFRI) has identified CLARIN ERIC (Common Language Resources and Technology Infrastructure) as one of two Research Infrastructures (RIs) of pan-European interest that meet the long-term needs across all scientific areas including social and cultural innovation. Speech technology has the potential to play a key role in cultural heritage, an area where speech is prevalent, but rarely approached with objective and efficient tools.

One of the projects currently running at KTH is a project that develops and makes available speech-to-text, or speech recognition, that is specifically adapted to work on archive materials. There are huge sets of speech and audio data in Swedish archives that cannot be used because they cannot be indexed, simply due to their size. The project produces speech technology research impact in the *Cultural heritage* area by making digitized audio materials truly available digitally, and pushes Swedish speech recognition forward by developing analysis and adaptation methods and by experimenting with new, previously unavailable training data. The application and tool, Edyson, presented in section 3 below, is an example of the innovative applications that are being developed at KTH in this area.

2.2 Inclusion & Accessibility

Inclusion and Accessibility is a key area for a sustainable and humane future society in a number of policies and strategies. Among these, we find UN Resolution 70/1, "Transforming our World: the 2030 Agenda for

Sustainable Development". The European commission's policy "*Digital Inclusion for a better EU society*" aims to ensure that everybody can contribute to and benefit from the digital economy and society through the development of assistive technologies. The Swedish government has published an action plan for agenda 2030, which again points to inclusion as a key area. Among its directives, it states that people's ability to participate in society shall not be governed by their background, their needs, or their preconditions. Speech technology plays a key role here, in particular with respect to the wide range of people who for whatever reason are excluded from written information.

A current project at KTH develops and makes available Swedish text-to-speech, or speech synthesis, that is suitable for reading aloud the kind of lengthy, complicated texts found in books. Talking books and audio books are notoriously expensive to produce, yet they are essential to inclusion. The project has a special focus on designing and testing innovative evaluation methods for measuring the usability of speech synthesis for the general public and for people with cognitive impairments. The ability to control speaking style, speech rate and articulatory clarity to optimally cater for different users and listening conditions is also a research topic. The project produces speech technology research impact in the *Inclusion and Accessibility* area by paving the way for making all books available as talking books or audio books. It pushes Swedish speech synthesis forward by researching methods specifically targeting the particular difficulties that arise when synthesizing read aloud books, as well as defining currently non-existing evaluation criteria for such speech syntheses.

2.3 Health & Aging

The Europeans Commission's policy "*Research and innovation in digital solutions for health, wellbeing and ageing*" includes both innovating health systems and promoting technology that supports healthy and independent living for the elderly. The range of successful speech technology applications in this area is growing at significant speed. Examples include support systems for the health sector, such as automatic transcription of prescriptions; teaching and training applications; and systems for diagnosis, prevention, treatment and rehabilitation.

The project, "Dialogue for Rehabilition," develops and makes available methods for human-computer dialogues designed to assist with (early) diagnosis, prevention, and rehabilitation. There are a number of health care areas in which relatively repetitive and simple dialogues are used for these purposes (e.g. dementia, autistic spectrum disorders, Parkinson). The project aims to research and generalize these dialogues and to create a dialogue platform that is specifically designed for their implementation. Integrity issues pose a major hurdle here, as they make it difficult, for example, to use commercial cloud based solutions. The project produces speech technology research impact in the *Health & Ageing* impact area by laying the foundation for a kind of patient-machine dialogue that has already proven very efficient in experiments. It pushes Swedish spoken dialogue system research by providing a new clear and useful dialogue type and connecting this to end users, and it will drive research into speech anonymization methods.

2.4 Digitalization

The area of digitalization is highlighted by the European Commission, by the Swedish government, and by KTH's long-term strategy as important for providing new revenue and value-producing opportunities in the process of moving to digital business and customer services. Conversational AI platforms using speech technology are seen to be among the strongest instigators of investments that exploit AI in the near future. This area of speech technology is one of the key strengths of Swedish speech technology in general and of KTH in particular.

At KTH we are building Conversational AI systems that give non-experts easy access to advanced support and guidance. In this project, we explore how conversational systems can be improved through better use of interactional data and by exploring additional modalities such as gaze and breathing. We focus on methods that detect human activities and affective states, and investigate how these can improve conversational skills of e.g. social robots and intelligent voice assistants. The project concerns the digitalization of companies and their processes with the help of speech technology.

3. Edyson

This section describes Edyson, a web-based framework for browsing and annotating large amounts of speech and audio data.

3.1 Temporally disassembled audio

Edyson is based on the notion of temporally disassembled audio (TDA), which is the idea of deconstructing an audio file along its temporal axis with the intent of producing a set of unordered sound snippets of short duration (Fallgren, Malisz, and Edlund, 2019a). Given a set of these short sounds one could rearrange them, and as such listen to them, in any order or manner one wants. Perhaps most importantly, the sounds do not need be ordered in a conventional 1-dimensional sequence, but could for instance be arranged along two axes according to some feature – as is the case of Edyson. The purpose of the process is to remove the time constraints that typically come with analyzing large audio files manually.

3.2 Audio processing

Given an audio file, the audio processing pipeline of Edyson conceptually consists of three main steps. First, the audio is temporally disassembled into snippets of equal length, typically in the range of a few hundred milliseconds to a couple seconds. Second, feature extraction is performed for every snippet, as such representing the short sound as a vector. The decision of what features to use is a tough problem in itself, and depends on the nature of the sound and the purpose of the analysis. MFCCs (Eyben, Wöllmer and Schuller, 2010) are, however, commonly used for speech and are as such used as default in Edyson. Third, the feature vectors are then run through a dimensionality reduction algorithm, e.g. t-SNE (Maaten and Hinton, 2008) self-organizing maps (Kohonen, 1982) that maps every vector to two dimensions - effectively generating a set of xy coordinates for every sound snippet.

3.3 Interface and functionality

The set of coordinates outputted by the audio processing pipeline are visualized in a 2D plot. The distribution of points, and potentially formed clusters, is based on the nature of the feature space, along with whatever properties of the sound the dimensionality reduction algorithm deems most prominent. In other words, two points that are similarly distributed in the plot should also have similar acoustic properties.

The Edyson interface (see Figure 1) has a list of functionalities of which only the essentials will be covered here; for a more extensive list see (Fallgren, Malisz and Edlund, 2019a) or the online documentation[1]. The most important functionality is the listening function, which allows the user to listen to the temporally disassembled audio. This is done by simply hovering over a region of points using the cursor, the system then samples randomly from the selected points and plays the sounds with some overlap which produces a blend of sounds. The parameters of the listening function can be adjusted in real-time in the Edyson interface. For more information on the listening function see Cocktail (Edlund, Gustafson and Beskow, 2010; Fallgren, Malisz and Edlund, 2018). If the user finds an interesting region they can assign a label to it by coloring the points of interest; the timeline then provides instantaneous feedback on where the colored region occurs in the original audio. Furthermore, the dimensionality reduction algorithm can be dynamically changed in real-time which may give the user new information. Edyson can also output any potential findings, specifically, the export function temporally reassembles every sound snippet with their respective label (color). The output can then be imported into other software for further analysis.

3.4 Exploration and annotation

The reason for using Edyson is at least twofold. First, it is an appropriate method for browsing some audio quickly and as such a way for researchers to gain insight into the nature of their data. This is a task that might seem trivial at first, but it is often challenging given the large size of modern audio collections. As an example, the National Library of Sweden in Stockholm hosts many millions of hours of audio-visual data. It is entirely conceivable that a lot, if not most, of these data, are not properly labeled, as is the case for other speech archives and audio collections. The process of TDA allows for fast and efficient browsing of audio which greatly facilitates many downstream tasks within research and audio analysis.

Edyson can also be used for annotation; however, it should be noted that its purpose is not to rival existing software highly specialized for control and efficiency for annotation. Rather, the annotation functionality in Edyson simply serves to provide the user with a basic set of labels of their findings, that for instance could be refined in further analysis.

3.5 Previous results and application areas

Although Edyson is still a work in progress it has shown potential for several different tasks. Fallgren, Malisz, and Edlund (2018) presented an early version and found evidence for the exploration aspect of the method. Specifically, it was shown that the TDA approach could be used to gain insight into different types of audio, e.g. speech, short speech segments, music, and animal sounds. Fallgren, Malisz and Edlund (2019b) conducted experiments where participants explored and properly labeled speech and applause segments in ~10 hours of presidential speeches in a matter of minutes. Most

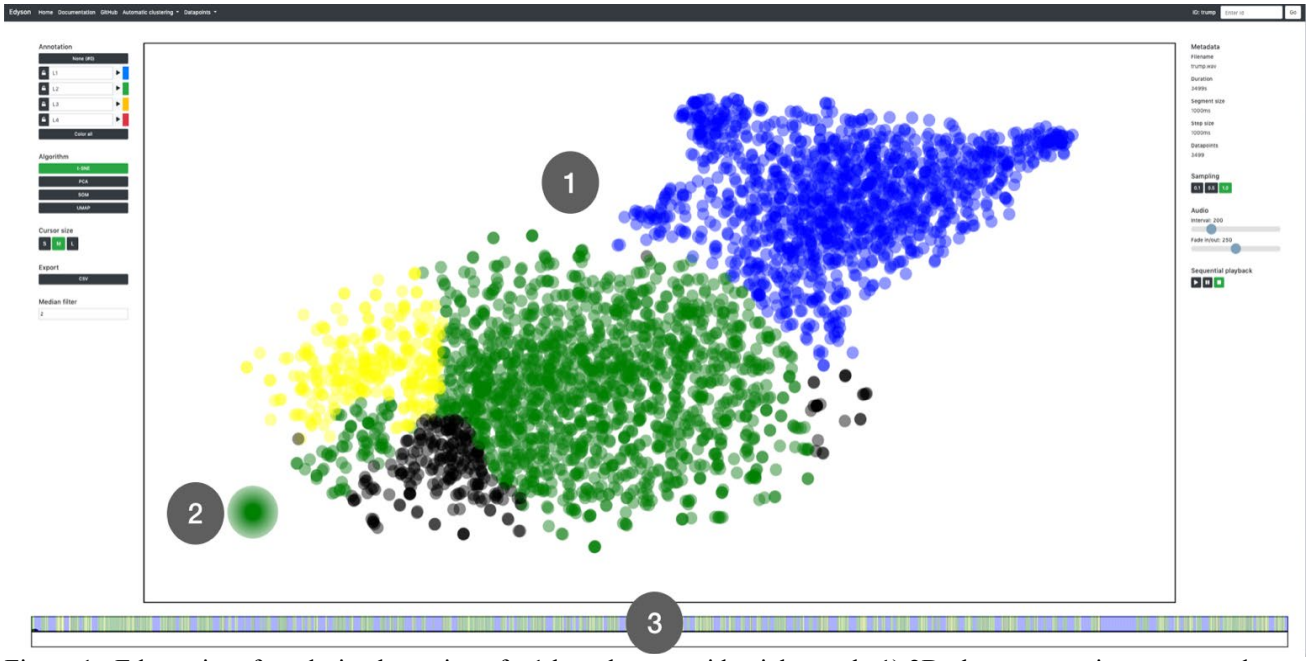

Figure 1 : Edyson interface during browsing of a 1 hour long presidential speech. 1) 2D plot, every point corresponds to a 1 second segment ; 2) Cursor, used for listening and coloring of points ; 3) Timeline, gives instant feedback when coloring.

[1] See github.com/perfall/Edyson for further information and installation instructions.

importantly, the participants did not have any prior knowledge of what the audio contained. Fallgren, Malisz, and Edlund (2019a) provided further evidence for the exploration and annotation aspect of Edyson by annotating 100 hours of noisy radio transmission data for speech activity in less than 45 minutes. Furthermore, pilot results have shown that the tool can be used to gain insights in several areas of speech and audio analysis, e.g. vowel detection, speaker separation, music browsing, noise detection to name a few. Potential application areas are mostly limited by the selection of features, as they carry the information that may or may not capture certain aspects of sound.

Currently Edyson is being used in the project TillTal (Berg et al. 2016), that aims to make cultural heritage recordings accessible for speech research. The Institute for Language and Folklore (ISOF) is engaged in the project and hosts more than 20,000 hours of speech recordings, most of which are digitized. The technology presented here has proven to be a fruitful resource regarding the task of utilizing the large quantities of speech data at hand.

Another reason for using Edyson is that it is completely language independent; one could even explore a recording containing two different spoken languages with the hope of finding similarities or distinctions. It may also help reveal contents of one's data that would otherwise not be found. For instance, when browsing the contents of an hour-long archived interview it was directly evident that there was a minute-long violin-segment in the middle of the recording. In many scenarios observations like this are important and shed light upon the importance of human-in-the-loop frameworks like Edyson.

4. Conclusion

There is currently a wide range of diverse activities in Sweden, particularly at KTH, for research and development of speech technology oriented to the Swedish language. While much of this activity is specifically oriented to Swedish, many of the resulting tools and applications, such as the tool, Edyson, presented here, can be used or adapted for use for any language.

5. Acknowledgements

The work reported on here is supported by the Swedish Research Council (Swe-Clarin and the National Swedish Language Bank, VR 2017-00626) and the Swedish Foundation for the Humanities and Social Sciences (SAF16-0917:1).

6. Bibliographical References

Berg, J., Domeij, R., Edlund, J., Eriksson, G., House, D., Malisz, Z., Nylund Skog, S. and Öqvist, J. (2016). Tilltal – making cultural heritage accessible for speech research. Proceedings CLARIN Annual Conference 26–28 October, Aix-en-Provence, France.

Fallgren, P., Malisz, Z., and Edlund, J. (2018). Bringing order to chaos: a non-sequential approach for browsing large sets of found audio data. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resource Association (ELRA).

Fallgren, P., Malisz, Z., and Edlund, J. (2019a). How to annotate 100 hours in 45 minutes. Proc. Interspeech 2019, pages 341-345.

Fallgren, P., Malisz, Z., and Edlund, J. (2019b). Towards fast browsing of found audio data: 11 presidents. In Proceedings of Digital Humanities in the Nordic Countries, DHN 2019, pages 133-142, Copenhagen, Denmark.

Edlund, J., Gustafson, J., and Beskow, J. (2010). Cocktail– a demonstration of massively multi-component audio environments for illustration and analysis. Proceedings Third Swedish Language Technology Conference (SLTC 2010) page 23, Linköping, Sweden.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, pages 1459-1462. ACM.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1): 59-69.

Maaten, L. V. D., and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research,* 9: 2579-2605.