Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Wilcock, Graham; Laxström, Niklas; Leinonen, Juho; Smit, Peter; Kurimo, Mikko; Jokinen, Kristiina

Towards SamiTalk: A Sami-Speaking Robot Linked to Sami Wikipedia

Published in:
Dialogues with Social Robots - Enablements, Analyses, and Evaluation

DOI:
10.1007/978-981-10-2585-3_28

Published: 01/01/2017

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Wilcock, G., Laxström, N., Leinonen, J., Smit, P., Kurimo, M., & Jokinen, K. (2017). Towards SamiTalk: A Sami-Speaking Robot Linked to Sami Wikipedia. In K. Jokinen, & G. Wilcock (Eds.), *Dialogues with Social Robots - Enablements, Analyses, and Evaluation* (pp. 343-351). (Lecture Notes in Electrical Engineering; Vol. 999). Springer. https://doi.org/10.1007/978-981-10-2585-3_28

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Towards SamiTalk: a Sami-speaking Robot linked to Sami Wikipedia

Graham Wilcock[1], Niklas Laxström[1], Juho Leinonen[2], Peter Smit[2],
Mikko Kurimo[2], and Kristiina Jokinen[1]

[1] University of Helsinki, Helsinki, Finland.
`firstname.lastname@helsinki.fi`
[2] Aalto University, Espoo, Finland.
`firstname.lastname@aalto.fi`

Abstract. We describe our work towards developing SamiTalk, a robot application for the North Sami language. With SamiTalk, users will hold spoken dialogues with a humanoid robot that speaks and recognizes North Sami. The robot will access information from the Sami Wikipedia, talk about requested topics using the Wikipedia texts, and make smooth topic shifts to related topics using the Wikipedia links. SamiTalk will be based on the existing WikiTalk system for Wikipedia-based spoken dialogues, with newly developed speech components for North Sami.

Keywords: language revitalisation · speech technology · humanoid robots · spoken dialogue systems

1 Introduction

In this paper we describe our work towards developing SamiTalk, a robot application for the North Sami language. This robot application is chosen because it is an interface to collaboratively edited Wikipedia information, and as a novel application, it is expected to increase the visibility of the language as well as interest in it. In particular, it is expected that young people may become more interested in using the language, which is regarded as an important and effective strategy for language revival in general. The motivation for creating robot applications to support revitalisation of endangered languages is discussed in more detail in [1].

SamiTalk will be the first robot application in the North Sami language, enabling users to hold spoken dialogues with a humanoid robot that speaks and recognizes North Sami. The robot will access information from the Sami Wikipedia, will talk about requested topics using the Wikipedia texts, and will make smooth topic shifts to related topics using the Wikipedia links. SamiTalk will be based on the existing WikiTalk system for Wikipedia-based spoken dialogues (see Section 2), with newly developed speech components for North Sami.

The paper is structured as follows. Section 2 summarizes the existing WikiTalk system and explains the differences between a localisation of WikiTalk for

North Sami and previous localisations. Section 3 describes the development of the speech synthesizer and speech recognizer for North Sami that will be essential components of SamiTalk. Section 4 gives an example of the style of interaction for getting Wikipedia information with Samitalk. Section 5 indicates plans for future work.

2 WikiTalk and SamiTalk

The WikiTalk system [2,3] accesses Wikipedia directly online. Using paragraphs and sentences from Wikipedia, the system can talk about whatever topics are of interest to the user. There are other applications that can read out Wikipedia articles, but simply reading out an article is a monologue rather than a dialogue. The key feature that enables WikiTalk to manage dialogues rather than monologues is its ability to handle smooth topic shifts, as described by [2]. Hyperlinks in the Wikipedia text are extracted, to be used as potential topic shifts. The system predicts that the user will often want to shift the topic to one of the extracted links.

The main challenge for dialogue modelling in WikiTalk is to present information in a way that makes the structure of the articles clear, and to distinguish between two conditions: the user shows interest and wants the system to continue on the current topic, or the user is not interested in the topic and the system should stop or find some other topic to talk about [3].

WikiTalk has been internationalised, and localised versions are now available for English, Finnish and Japanese. A number of major issues in internationalisation and localisation of spoken dialogue systems are discussed by [4], using WikiTalk as an example. In the case of WikiTalk, each new language version requires speech recognition and speech synthesis components for the language, an available Wikipedia in the language, and a localisation of the WikiTalk multimodal user interface.

Due to the internationalisation of WikiTalk, new language localisations can generally be produced relatively rapidly. However, a localisation of WikiTalk for North Sami involves more than previous localisations of WikiTalk. Up to now, the speech recognition and speech synthesis components for the languages supported by WikiTalk have been the components provided by the Nao robot. In order to use these languages, the WikiTalk system simply needs to check that the required language is installed on the robot. All Nao robots can speak English, but other languages need to be installed. Currently 19 languages (including Finnish and Japanese) are available for purchase. If the required language is installed on the robot, it can be selected and set as the current language until another language is selected.

The situation for North Sami is quite different, as it is not one of the 19 available languages. The required speech recognition and speech synthesis components are being newly developed for the DigiSami project using the new DigiSami corpus.

3 Speech technology for SamiTalk

Both a speech recognizer and speech synthesizer are essential parts for a Sami-speaking robot. Ideally the robot would be able to speak any free-form North Sami sentence and be able to recognize any spoken sentence from any North Sami speaker. Given the amount of available resources however, this recognizer is not attainable at the moment and the current recognizer can either understand large vocabulary continuous speech from a single speaker, or only a small list of words, e.g. a list of common Wikipedia titles, from a wide range of speakers.

The biggest obstacle to creating these systems is the limited availability of high quality data. Especially for large-vocabulary continuous speech recognition, which normally requires more then 100 different speakers for a good model, the amount of available data limits both functionality and performance.

In this research we used audio data available from only two speakers; one male and one female. The speech is clean, read speech and amounts to 4.6 and 3.3 hours respectively. Two text corpora are available. The first is a download of the North Sami Wikipedia which contains 10K sentences with 20K different word types and the second is *Den samiske tekstbanken* provided by the University of Tromsø, which has almost 1M sentences with 475K different word types.

3.1 Speech Synthesis

A statistical Text-To-Speech (TTS) system was built using the Ossian toolkit[3], which is suitable for building synthetic voices from small amounts of data with a minimal requirement of linguistic expertise.

A TTS system commonly has two parts. The first part predicts a sequence of labels from a piece of plain text. These labels contain a multitude of information, such as the phoneme, the phoneme context, stress and prosody information. Normally these labels are generated using rule-based systems and lexicons that were purposely designed for the language. The Ossian toolkit however, utilizes vector space models to predict the pronunciation, stress and other relevant factors needed to create a voice [5]. Hence, the process does not require any linguistic expertise or resources.

The second part of the TTS systems uses the generated labels to create an audio file. There are two common approaches, Statistical Parametric Speech Synthesis (SPSS) and Unit Selection Systems (USS). SPSS systems require in general less data than USS systems, but generate a less natural result. As the data is limited, we created an SPSS model, which uses the HMM-based Speech Synthesis System (HTS) and the GlottHMM vocoder [6].

Given the small amount of data used (approximately 3 hours), the system performs reasonably well and informal listening tests show that the results are approaching those of a commercial solution by Acapela. The main difference is that the commercial system has a hand-crafted preprocessor to transcribe numbers and abbreviations and the use of punctuation to determine prosodic structures.

---

[3] Open source, available from `http://simple4all.org/product/ossian/`

3.2 Speech Recognition

An Automatic Speech Recognition (ASR) system has two main components, the acoustic model and the language model, trained from the audio data and the text data respectively. In order to make an ASR system that can recognize a large vocabulary continuous speech of any person, both the acoustic model and the language model need to be big enough and of high quality. Unfortunately a high quality acoustic model that is speaker-independent - i.e. it can perform recognition with any speaker of a language, not only speakers present in the training data - normally requires data from at least 100 different speakers. As this data is not available, we designed two other recognizers. The first recognizes a large vocabulary but is speaker-dependent - i.e. it can perform recognition with only one specific speaker. The second recognizes with any speaker, but with a limited vocabulary.

The acoustic model of the first recognizer, the speaker-dependent system, is trained using conventional methods using the data of only one speaker. The amount of audio data available is enough to build a well-performing system [7]. The language model is trained in a similar way as is done normally for Finnish, using a sub-word language model [8] to combat the high number of word forms in the North Sami languages that are caused by its agglutinative nature.

The second, speaker-independent, model is trained in a more unconventional way. Instead of using North Sami speech data, a big database of Finnish speech with over 200 different speakers is used. After that, the sounds (or phonemes) of North Sami are mapped to the closest sounding Finnish phoneme. This gives a speaker-independent system which has limited accuracy because of the initial language mismatch. To overcome the limited accuracy of the acoustic model, the language model is dramatically reduced in complexity, to a small vocabulary system. The exact vocabulary can be varied based upon the needs of the dialogue task, but it should stay small to keep an acceptable performance. This system requires also a significant amount of linguistic knowledge as the mapping between Finnish and North Sami words has to be done by hand.

Both systems are tri-state tri-phone hidden Markov models with Gaussian mixture model emission distributions, trained with the AaltoASR[4] toolkit [9,10]. In the case of the large vocabulary system, Morfessor[5] [11] is used to split words into segments to reduce the number of types in the lexicon. For language modelling a varigram model created with VariKN toolkit[6] [12] is used. The language model is trained on the *Den samiske tekstbanken* corpus.

In [7] the performance of the large vocabulary speaker-dependent system is compared with similar systems in Finnish and Estonian. The reported accuracy is similar to the systems in other languages and gives on average a 23% Word Error Rate (WER) for a speaker-dependent system trained with 2.5 hours of data. In the other language systems the error rate can be decreased to approximately 18% WER if 8 hours of data per speaker is available. In [7] it is also shown that

---

[4] Open source, available from `https://github.com/aalto-speech/AaltoASR`

[5] Open source, available from `http://www.cis.hut.fi/projects/morpho/`

[6] Open source, available from `https://github.com/vsiivola/variKN`

if a full speech database with more than 100 different speakers were available, it would be possible to build a large vocabulary speaker-independent system with an expected WER around 20%.

Overall we have succeeded in making a synthesizer and recognizer which are suitable for the SamiTalk application. Further development requires either more data or new techniques to create high quality speaker-independent models, possibly using the more widely available data of related languages such as Finnish.

4 Example interaction

This section shows an example of how users will interact with SamiTalk. The robot suggests a few topics that it can talk about using information from Sami Wikipedia. Among them, *Tiina Sanila-Aikio* is the current president of the Sami Parliament of Finland, and Japanese *Manga* shows that "local" topics such as *Snowmobile* are not the only topics in Sami Wikipedia. The user asks the robot to talk about the situation of the Sami language in Finland.

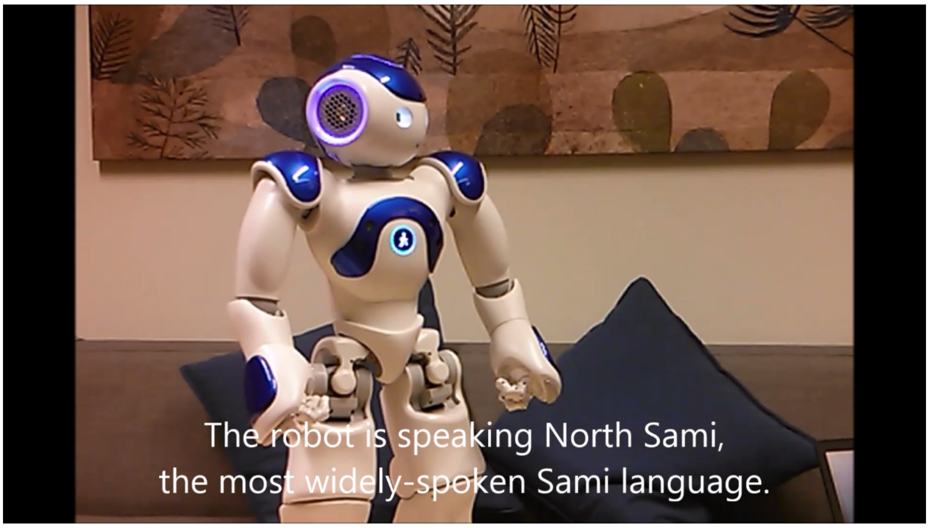
The robot is speaking North Sami,
the most widely-spoken Sami language.

Fig. 1. Screenshot from an example interaction with SamiTalk on a Nao robot. The video can be viewed at `http://www.helsinki.fi/digisami/#samitalk`.

- Robot: *Man fáttá birra don háliidat gullat? Soames mu favorihta fáttát leat Sámegiella, Roavvenjárga, Mohtorgielká, Tiina Sanila-Aikio, Manga.* (What topic do you want to hear about? Some of my favourites are: Sami language, Rovaniemi, Snowmobile, Tiina Sanila-Aikio, Manga.)
- User: *Sámegiella* (Sami language)

- Robot: *Sámegiela dilli Norgga, Ruoŧa, Suoma ja Ruošša bealde lea earálágan.* (The situation of Sami language is different in Norway, Sweden, Finland and Russia.)
- User: *Dilli Suomas* (The situation in Finland.)
- Robot: *Bealli Suoma 9 350 sámis máhtet sámegiela. Suomas hállojuvvo davvisámegiella, anárašgiella ja nuortalašgiella, main anárašgiella dušše Suomas. Buoremus dilli lea davvisámegielain, mii ii leat seamma áitatvuloš go eará sámegielat. Anárašgiela dilli lea vehá buoret go nuortalašgiela, dan dihte go anárašgielat giellabeassi doaibma lea joatkašuvvan guhkit áiggi. Suomas sámiid ruoktoguovllus ássi sámiin lea leamaš riekti geavahit sámegiela virgeolbmuiguin jagi 1992 rájes, muhto virgeolbmuin ii gáibiduvvo sámegielmáhttu baicce geavahuvvojit jorgaleaddjit ja dulkat.* (Half of the 9,350 Sami in Finland speak Sami language. Three Sami languages are spoken in Finland: North Sami, Inari Sami and Skolt Sami. Inari Sami is spoken only in Finland. North Sami has the best situation, and it is not as endangered as other Sami languages. The situation of Inari Sami is a bit better than of Skolt Sami, because the language nest of Inari Sami has existed longer than the language nest of Skolt Sami. In Finland the Sami people who live in the Sami region have had the right to use Sami language with officials since 1992, but the language knowledge is not required of officials, they use translators and interpreters.)
- Robot: *Leage buorre, vállje čuovvovaččain: Joatkke, Ovddit, dahje daja Bisán nu ahte bisánuvvo.* (Please select from: Continue, Previous, or say Stop if you want to stop.)
- User: *Joatkke.* (Continue.)
- Robot: *Suoma sámiid ruovttoguovllus mánáin lea riekti lohkat vuođđoskuvllas ja logahagas sámegiela eatnigiellan, válljaávnnasin dahje eaktodáhtolaččat. Dasa lassin giella oahppahuvvo maid ruovttoguovllu olggobealde Oulus, Roavvenjárggas ja Soađegili girkogilis. Sámegiela lea vejolaš čállit studeantadutkosis eatnigiellan. Vuosttas háve dát dáhpáhuvai jagis 1994.* (In the Sami region in Finland the children have the right to study Sami as mother tongue, optional subject, or voluntarily in basic school and senior secondary school. In addition, the Sami language is taught outside the Sami region, in Oulu, Rovaniemi and the center of Sodankylä. It is possible to take the mother tongue part of the matriculation examination in Sami. This was done for the first time in 1994.)
- User: *Joatkke.* (Continue.)
- Robot: *Sámit leat ain juo guovttegiellagat, mii dárkkuha dan, ahte sii hállet sámegiela lassin riikkaid váldogiela. Suomas skuvllain oahpahuvvo davvisámegiela lassin maid anáraš- ja nuortalašgillii, vaikke eanas oahpahusas lea davvisámegillii. Sámegielat oahpahusa birrasis leat Suomas sullii 500 oahppi, geain 150 oźźot oahpahusa sámegillii. Váldooassi oahpahusas dáhpáhuvvá 1-6-luohkáin.* (The Sami are now always bilingual, which means that in addition to Sami they speak the main language of the state they live in. In Finland the schools teach in North Sami, Inari Sami and Skolt Sami, although mostly the teaching language is North Sami. Under the teaching in

Sami language there are around 500 pupils in Finland, and 150 of them gain teaching in Sami. The main part of the teaching in Sami language is given in classes 1-6.)
- User: *Joatkke.* (Continue.)
- Robot: *Suomas golbma universitehta fállet sámegiela oahpahusa. Oulu universitehta Giellagas-instituhtas lea vejolaš studeret sámegiela váldoávnnasin. Oalgeávnnasin sámegiela lea vejolaš studeret Helssega universitehtas ja Lappi universitehtas.* (In Finland three universities offer education in Sami language. The Giellagas Institute of Oulu University offers Sami language as main subject. It is possible to study Sami language as minor subject at University of Helsinki and University of Lapland.)
- User: *Bisán.* (Stop.)
- Robot: *Háliidatgo duođat bisánit?* (Do you really want to stop?)
- User: *Juo.* (Yes.)
- Robot: *Ortnegis. Giitu. Báze dearvan.* (OK. Thank you. Goodbye.)

5 Conclusions and future work

The speech components for North Sami are under further development and have not yet been integrated and tested with the Nao robot software and the WikiTalk software. For demo purposes we hope to use a speaker-independent, but closed vocabulary, speech recognizer. The dialogue has been restricted in order to show the general concept even with this limitation.

In future work, we will explore options for creating a large-vocabulary continuous speech recognizer that is speaker-independent with an open vocabulary. The options include not only collecting more data, but also techniques for utilizing available North Sami resources such as news broadcasts, and resources from related Finno-Ugric languages such as Finnish.

Acknowledgements

We thank our DigiSami project colleagues for Sami and English translations and the University of Tromsø for access to their North Sami datasets, and acknowledge the computational resources provided by the Aalto Science-IT project. The work has received funding from the EC's 7th Framework Programme (FP7/2007-2013) under grant agreement n°287678 and from the Academy of Finland under the Finnish Centre of Excellence Program 2012–2017 (grant n°251170, COIN) and through the project Fenno-Ugric Digital Citizens (grant n°270082).

References

1. Jokinen, K., Hiovain, K., Laxström, N., Rauhala, I., Wilcock, G.: DigiSami and digital natives: Interaction technology for the North Sami language. In: Proceedings of Seventh International Workshop on Spoken Dialogue Systems (IWSDS 2016). Saariselkä (2016)

2. Wilcock, G.: WikiTalk: A spoken Wikipedia-based open-domain knowledge access system. In: Proceedings of the COLING 2012 Workshop on Question Answering for Complex Domains. pp. 57–69. Mumbai (2012)
3. Jokinen, K., Wilcock, G.: Multimodal open-domain conversations with the Nao robot. In: Mariani, J., Rosset, S., Garnier-Rizet, M., Devillers, L. (eds.) Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialogue Systems into Practice, pp. 213–224. Springer (2014)
4. Laxström, N., Wilcock, G., Jokinen, K.: Internationalisation and localisation of spoken dialogue systems. In: Proceedings of Seventh International Workshop on Spoken Dialogue Systems (IWSDS 2016). Saariselkä (2016)
5. Watts, O.: Unsupervised Learning for Text-to-Speech Synthesis. Ph.D. thesis, University of Edinburgh (2012)
6. Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., Alku, P.: HMM-based speech synthesis utilizing glottal inverse filtering. IEEE Transactions on Audio, Speech, and Language Processing 19(1), 153–165 (2011)
7. Smit, P., Leinonen, J., Jokinen, K., Kurimo, M.: Automatic speech recognition for Northern Sami with comparison to other Uralic languages. Presentation at Second International Workshop on Computational Linguistics for Uralic Languages, Szeged, Hungary (2016)
8. Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J.: Unlimited vocabulary speech recognition with morph language models applied to Finnish. Computer Speech and Language 20(4), 515–541 (2006)
9. Hirsimäki, T., Pylkkönen, J., Kurimo, M.: Importance of high-order n-gram models in morph-based speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 17(4), 724–732 (2009)
10. Pylkkönen, J.: An efficient one-pass decoder for Finnish large vocabulary continuous speech recognition. In: Proceedings of the 2nd Baltic Conference on Human Language Technologies. pp. 167–172 (2005)
11. Virpioja, S., Smit, P., Grönroos, S.A., Kurimo, M.: Morfessor 2.0: Python implementation and extensions for Morfessor baseline. Tech. rep. (2013)
12. Siivola, V., Creutz, M., Kurimo, M.: Morfessor and variKN machine learning tools for speech and language technology. In: INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association. pp. 1549–1552 (2007)