Chapter 15
Basque-speaking Smart Speaker based on Mycroft AI

Igor Leturia, Ander Corral, Xabier Sarasola, Beñat Jimenez, Silvia Portela, Arkaitz Anza, and Jaione Martinez

Abstract Speech-driven virtual assistants, known as smart speakers, such as Amazon Echo and Google Home, are increasingly used. However, commercial smart speakers only support a handful of languages. Even languages for which ASR and TTS technology is available, such as many official EU member state languages, are not supported due to a commercial disinterest derived from their – relatively speaking – rather small number of speakers. This problem is even more crucial for minority languages, for which smart speakers are not expected anytime soon, or ever. In this ELG pilot project we developed a Basque-speaking smart speaker, making use of the open source smart speaker project Mycroft AI and Elhuyar Foundation's speech technologies for Basque. Apart from getting it to speak Basque, one of our goals was to make the smart speaker privacy friendly, non-gendered and use local services, because these are usual issues of concern. The project has also served to improve the state of the art of Basque ASR and TTS technology.

1 Overview and Objectives of the Pilot Project

Commercial smart speakers are increasingly popular despite the fact that their language coverage leaves much to be desired. Many large official national languages and practically all minority languages are unsupported by these devices. In many cases, the lack of support for a language in a smart speaker is not due to the lack of the necessary speech technologies, i. e., Automatic Speech Recognition (ASR) and Text To Speech (TTS). ASR and TTS technologies do exist for the Basque language

Igor Leturia · Ander Corral · Xabier Sarasola
Elhuyar Fundazioa, Spain, i.leturia@elhuyar.eus, a.corral@elhuyar.eus, x.sarasola@elhuyar.eus

Beñat Jimenez
Talaios Koop., Spain, jimakker@talaios.coop

Silvia Portela · Arkaitz Anza · Jaione Martinez
Skura Mobile, Spain, silvia@skuramobile.com, arkaitz@skuramobile.com, jaione@skuramobile.com

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_15

but it is unlikely that they will be implemented in smart speakers developed by the big technology enterprises because of its relatively small number of speakers.

On the other hand, there is a rather mature, open source smart speaker project called Mycroft AI.[1] Our ELG pilot project develops an open source smart speaker for the Basque language, based on Mycroft AI, that makes use of Elhuyar Foundation's ASR and TTS technologies. Apart from being open source and in Basque, other points of interest were the handling of privacy, gender and service locality issues.

One objective of the project was to improve the state of the art of Basque ASR and TTS technologies, since it would be necessary to adapt them to the context of a smart speaker. Specifically, we wanted to 1. improve the performance of Basque ASR technology for noisy environments; 2. create a grammar-based ASR system instead of a general vocabulary one to only recognise the commands of the speaker and, thus, improve precision; 3. create a neural network-based TTS system for Basque and replace the old HMM one; and 4. try to develop a gender-neutral voice.

2 Mycroft Localisation

A crucial and necessary part of the project was the localisation of Mycroft to Basque in its broadest sense. This involved not only a string translation process, but also making it understand speech commands and respond via speech in Basque. Thus, we had to develop plugins to connect Mycroft to Elhuyar's ASR and TTS services.

The localisation also involved the adaptation to Basque of Mycroft's linguistic module called *lingua-franca*, responsible for parsing numbers, days, times, durations, etc. in speech commands and to pronounce them correctly when responding.

Finally, the routine job of string translation of any software localisation process turned out not to be as straightforward for the commands' part. The parsing of many skills' intents from the commands is done by simply detecting some required or optional keywords and parameters, which is why their translation required more than just a simple sentence translation. We translated the Mycroft core module and 40+ of its skills (volume control, date, time, lists, alarms, audio record, radio, news, Wikipedia, weather, jokes, Wikiquote, e-mail etc.).

3 Privacy, Gender and Proximity

As mentioned in Section 1, we wanted to address the privacy and gender concerns often associated with smart speakers and also promote the use of local services. Regarding privacy, users and potential buyers have concerns with having a device in their homes with a microphone that is always on (Lau et al. 2018). However, respect for privacy is precisely one of Mycroft AI's unique selling propositions. They claim

[1] https://mycroft.ai

that they are “private by default” and that they “promise to never sell your data or give you advertisements” using their technology. This materialises in the fact that the wake word (“Hey, Mycroft”) is detected locally, i. e., no audio is sent to remote servers except when saying a command after the detection of the wake word. On the other hand, if some big enterprise’s cloud-based ASR or TTS services are used for the recognition of commands and the utterance of responses, there are logically some doubts as to what these companies will do with that data. Using Elhuyar’s Basque ASR and TTS remote APIs from Mycroft, no data would be kept or collected.

Regarding gender treatment, smart speakers are known for their improper gender treatment, as stated in the Unesco report “I’d blush if I could: closing gender divides in digital skills through education” (West et al. 2019). According to this report, practically all commercial smart speakers exhibit a female voice and female personalities, and respond obligingly even to hostile requests, verbal abuse and sexual harassment, which may lead to reinforce and spread gender biases. The report ends with some recommendations that range from not making digital assistants female by default to developing neutral voices and personalities, which our project has tried to follow. The Basque voice installed at the moment is a male voice by default. Also, the speaker’s name, Mycroft, – although fictional – is male, its “personality” is neutral, and it has no skill to respond in a docile manner to sexual comments or verbal abuse. However, we have also carried out some experiments in order to develop a gender-neutral synthetic voice (see Section 4.4).

We felt that our smart speaker should prioritise the local region and, for instance, allow listening to local radio stations, read the news from local media or buy goods or order food from local stores. We developed half a dozen local skills of our own, including local news, local radio stations, dictionary querying or Basque music.

4 Developments in Basque Speech Technology

4.1 ASR Robustness in Noisy Environments

One of the main challenges regarding the use of ASR technology in a smart speaker is making it robust enough to be reliable under non-optimal conditions: low volume, background noise, music, speech, room reverberation, low quality microphone, etc.

Elhuyar’s ASR system for the Basque language is a general purpose system based on the Kaldi[2] toolkit. The speech data used to train the acoustic model comprises high quality clean parliamentary speeches. To make our acoustic model more robust, we used several synthetic data augmentation techniques during the training phase (Alumäe et al. 2018). This means that training data was 1. synthetically augmented by adding background noises from the MUSAN dataset (Snyder et al. 2015), which comprises several recordings of music, speech and a wide variety of noises;

[2] https://kaldi-asr.org

2. artificially reverberated with various real and simulated room impulse responses (Ko et al. 2017); and 3. augmented with threefold speed and volume perturbations.

4.2 ASR Closed Grammar-based Recognition

For general purpose ASR systems, typically a large language model is trained with a vast amount of diverse texts. For a smart speaker, however, where the user is expected to use a closed set of commands, limiting the ASR's vocabulary to just the necessary commands can increase the precision of the speech recognition.

Since Kaldi internally uses weighted finite state transducers (WFST) to model the language, simply by converting all the commands defined in Mycroft skills to the format used by Pynini (a Python library for WFST grammar compilation), we would obtain a language model limited to Mycroft's commands. But although Mycroft's skills were originally defined using its old-style intent parser Padatious (where the whole command is defined), nowadays most skills use the new intent parser Adapt, which defines commands using a few keywords and parameters. This makes it unfeasible to automatically generate all possible commands containing the keywords and parameters. Rewriting all skills to the Padatious format would have made the code much more difficult to maintain as well as losing Adapt's recall gain. This is why the creation of a custom grammar was eventually discarded.

4.3 Neural Network-based Basque TTS

Elhuyar's previous Basque TTS service was based on Hidden Markov Models (HMMs). In the ELG pilot project we developed a new neural network-based TTS service. Since the first neural system was published in 2013 (Zen et al. 2013), these have taken a clear advantage over HMM-based approaches and systems like Tacotron 2 (Shen et al. 2018) have achieved naturalness comparable to natural voice.

The key challenge with neural TTS systems is the size of the training dataset. The original Tacotron 2 monospeaker system was trained with 24.6 hours of speech, and subsequent research concluded that 10 hours is the minimum time required to obtain maximum quality (Chung et al. 2019). The only publicly available database of Basque speech of that size is a multispeaker database created by Google (Kjartansson et al. 2020), which contains recordings from 53 speakers with a maximum of 15 minutes per speaker. Modified configurations of Tacotron 2 using speaker embeddings have proved successful providing good quality multispeaker TTS systems (Jia et al. 2018), i. e., systems trained using combined recordings of multiple speakers, capable to synthesise the voice of each of them. We recorded a small multispeaker database, combined it with the Google database, and trained a multispeaker TTS using speaker embeddings, obtaining our own neural quality TTS voices.

4.4 Gender-neutral Voice

Apart from the interventions to address gender issues (Section 3), we conducted experiments towards obtaining a gender-neutral voice. Tolmeijer et al. (2021) observed that we do not regard voices of intermediate pitch (which is what could be understood as gender-neutral) as genderless, that we assign them one gender or the other, and that those that could be best considered as ambiguous in terms of gender or genderless were those with the greatest division of opinion.

Most of the literature on the field of generating gender-ambiguous voices seek gender neutrality through pitch modification, such as Tolmeijer et al. (2021), or the first genderless voice Q (Carpenter 2019). We employed a different and innovative approach. We first calculate the average speaker embedding for each gender with the embeddings obtained in the training and then we compute the embedding that is midway between the average male and female embeddings. Using this embedding in the trained Tacotron 2, we can synthesise sentences with a voice which has produced divided opinions as to its gender and which can thus be considered genderless.

5 Conclusions and Results of the Pilot Project

This ELG pilot project developed an open source Basque-speaking smart speaker based on Mycroft AI, which respects privacy and which uses a more appropriate approach regarding the voice's gender than commercial smart speakers. We connected Mycroft to Elhuyar's Basque ASR and TTS services, and we improved the state of the art of Basque speech technologies. Our ASR for Basque performs better in noisy environments and we developed a new deep neural network-based TTS for Basque and made experiments towards a gender-ambiguous synthetic voice. We translated more than 40 Mycroft skills and developed half a dozen new ones addressing local services. We tested the Basque Mycroft in PCs and Google AIY Kits.

Anyone can now download, install on a device and try Mycroft in Basque. While the ELG pilot project is finished, we continue to work on the project with the aim of, if possible, bringing a Basque smart speaker device to the market. We believe that the work carried out, the experience gained and the code developed in the ELG pilot project can be very useful for other minority language communities that would like to have access to a smart speaker that speaks their own language.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects. The project has also been helped by the Basque Government through its Hazitek programme (project DomEus) and the Gipuzkoa Provincial Council through its Etorkizuna eraikiz programme (project Mycroft.eus).

References

Alumäe, Tanel, Ottokar Tilk, and Asad Ullah (2018). "Advanced rich transcription system for Estonian speech". In: *Human Language Technologies – the Baltic Perspective: Proc. of the Eighth Int. Conference (Baltic HLT 2018)*. Ed. by Kadri Muischnek and Kaili Müürisep. Amsterdam, the Netherlands: IOS Press, pp. 1–8. DOI: 10.3233/978-1-61499-912-6-1.

Carpenter, Julie (2019). "Why Project Q is More than the World's First Nonbinary Voice for Technology". In: *Interactions* 26.6, pp. 56–59. DOI: 10.1145/3358912.

Chung, Yu-An, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan (2019). "Semi-supervised training for improving data efficiency in end-to-end speech synthesis". In: *ICASSP 2019*. IEEE, pp. 6940–6944.

Jia, Ye, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al. (2018). "Transfer learning from speaker verification to multispeaker text-to-speech synthesis". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, pp. 4485–4495.

Kjartansson, Oddur, Alexander Gutkin, Alena Butryna, Isin Demirsahin, and Clara E. Rivera (2020). "Open-Source High Quality Speech Datasets for Basque, Catalan and Galician". In: *SLTU-CCURL 2020*. 11–12 May, Marseille, France, pp. 21–27.

Ko, Tom, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur (2017). "A study on data augmentation of reverberant speech for robust speech recognition". In: *ICASSP 2017*, pp. 5220–5224. DOI: 10.1109/ICASSP.2017.7953152.

Lau, Josephine, Benjamin Zimmerman, and Florian Schaub (2018). "Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers". In: *Proc. of Human-Computer Interaction* 2.CSCW, pp. 1–31. DOI: 10.1145/3274371.

Shen, Jonathan, Ruoming Pang, Ron Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. (2018). "Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions". In: *ICASSP 2018*. IEEE, pp. 4779–4783.

Snyder, David, Guoguo Chen, and Daniel Povey (2015). "Musan: A music, speech, and noise corpus". In: *arXiv preprint arXiv:1510.08484*.

Tolmeijer, Suzanne, Naim Zierau, Andreas Janson, Jalil Sebastian Wahdatehagh, Jan Marco Marco Leimeister, and Abraham Bernstein (2021). "Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution". In: *Conference on Human Factors in Computing Systems (CHI)*. New York, NY, USA: ACM, pp. 1–7.

West, Mark, Rebecca Kraut, and Han Ei Chew (2019). *I'd blush if I could: closing gender divides in digital skills through education*. Unesco EQUALS.

Zen, Heiga, Andrew Senior, and Mike Schuster (2013). "Statistical parametric speech synthesis using deep neural networks". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 7962–7966.