INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Whistle-to-text: Automatic recognition of the Silbo Gomero whistled language

Agata Jakubiak

University of Warsaw, Poland
jakubiakagata42@gmail.com

Abstract

Automatic speech recognition (ASR) is a rapidly developing field of study. However, ASR for other types of speech than the regular spoken speech—for example, whispering or shouting—remains difficult, as it requires specific models trained to recognise these types of speech.

A lesser-known type of speech than those is the whistled speech, in which speech is transformed into whistling.

In this paper, I will describe how I created the first-ever ASR model designed to recognise a whistled language. It was trained, using the HMM-GMM approach to ASR, to recognise the whistled dialect of Spanish, Silbo Gomero.

This model learned to recognise Silbo Gomero, though its performance was somewhat worse than that of spoken speech recognition models trained on data sets of similar size. It appears that methods used to create spoken language ASR models can be used to create whistled language ASR models, with only small changes—which will be explained in this paper—required.

Index Terms: speech recognition, whistled speech, silbo gomero, whistled language, automatic whistled speech recognition

1. Introduction

Automatic speech recognition (ASR) is a broad and well-developed field of study. However, it has proven difficult to implement ASR for other methods of speaking than the regular spoken speech. ASR of whispered speech [1], shouted speech [2], and sung speech [3] usually requires a specialised ASR model, since regular ones fail to recognise those speech types correctly. Those specialised models, however, are created by the same methods used for spoken speech ASR; the main difference are the data sets used to train them, which contain the type of speech the model is being trained to recognise.

A relatively uncommon type of speech is the whistled speech, in which speaking is replaced with whistling. This is a more significant departure from speaking than those that occur in the previously mentioned whispering, shouting, and singing; an untrained fluent user of a language will not understand its whistled form. For this reason, communities using this kind of speech are said to use whistled languages. However, it is still a type of speech production, and not a language onto itself.

As of the time of writing, no research into ASR of whistled languages has been published. Because of that, I decided to study this subject, and determine whether one could create an Automatic Whistled Speech Recognition (AWSR) model with contemporary techniques used in creation of ASR models. This research focused on the whistled form of Spanish used on the La Gomera island of the Canary Islands archipelago—Silbo Gomero—as it is the most well-studied of the whistled languages, and it was possible to obtain enough recordings of it to conduct the experiments.

The ASR models necessary for this research were created using the Kaldi speech recognition toolkit [4], which utilizes the Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) approach to ASR. Kaldi was chosen because it is well studied, well documented, and highly customisable. Most modern ASR models utilize the Deep Neural Network (DNN) approach instead, but nothing about the methods used is specific to a HMM-GMM model, so they should be transferable.

2. Whistled Languages

Whistled languages are rare, but they are used in many places around the world, most often for long-distance communication in remote areas. Because of the isolation inherent in the practice, whistled languages remain relatively poorly documented.

2.1. Types of Whistled Languages

Whistled languages can be broadly separated into tonal whistled languages, and non-tonal whistled languages, based on the tonality of the language they're emulating.

Tonal whistled languages are based on tonal languages; a user of such a language whistles the tones that the message they're conveying would have when spoken, and omits other elements of speech [5, pp.122-123]. In highly tonal languages, this is enough to convey complex messages. A user of such a language might use drums or other musical instruments instead of whistling to transmit a message over large distances [6].

Non-tonal whistled languages are rarer and more complex; a user of such a language articulates with their mouth as they would when speaking, while also continuously whistling. The movement of their mouth—corresponding to the phonemes they would be speaking—changes the frequency and volume of their whistle, and a trained user can recognise the message transmitted in the sound, regardless of its complexity [7]. Since this form of speech is highly dependant on articulation, musical instruments can't be used to create the sound; instead, finger-whistling techniques are used for long-distance communication.

In this paper, ASR of non-tonal whistled languages will be explored, as Silbo Gomero falls into that category. ASR of tonal whistled languages would likely require different methods than described here.

2.2. Phonetics of Silbo Gomero

In Silbo Gomero, as in other non-tonal whistled languages, the sound of the whistle encodes the phonemes of spoken speech

10.21437/Interspeech.2023-989

Table 1: *Clusters of Spanish consonant sounds which are represented by single Silbo Gomero consonant sounds. Standard IPA notation is used.*

Phonemes represented by Silbo Gomero consonant sounds
[p],[k]
[b],[g]
[s]
[t],[d]
[n],[ɲ],[l],[r],[j],[ʝ],[ð],[ʎ]
[m]
[w]
[tʃ]
[f],[x]

[5, Chapter 7.2]. Every phoneme of spoken speech translates to a specific whistling sound, with vowels being transformed into periods of consistent frequency, and consonants being transformed into periods of changing frequency. Through that process, some consonants become indistinguishable from each other, as the shapes of the mouth used to speak them result in the same whistling sound; vowels, however, are usually easy to tell apart. Because of that simplification, languages with emphasis on vowels and relatively simple consonant systems—such as Spanish [8]—are better suited for transformation into whistled languages.

According to the phonetic analysis conducted by Classe [7], Silbo Gomero has nine consonant sounds. The clusters of consonants they represent are shown in Table 1.

Additionally, Silbo Gomero has five vowel sounds, corresponding to the five vowels of spoken Spanish: **i, e, a, o, u**. Since vowels provide about 50% of linguistic information in Spanish [7], the fact that they can be reliably recognised in their whistled form is crucial for intelligibility of whistled Spanish.

This simplified phonetic system retains enough complexity that fluent users of Silbo Gomero can communicate freely in it. Through the simplification of consonants, some words become indistinguishable from each other, but context is usually enough to resolve such situations.

2.3. Uses of Whistled Languages

Whistled languages are most commonly used for long-distance communication, typically in mountainous terrains, where the travel is difficult but the sounds spread well. Many whistled languages seem to have been developed by shepherds working in the mountains, who wished to talk to each other, but were often separated by large distances [5, Chapter 3.2.2]. By whistling with their fingers, a fluent user of a whistled language can emit sounds with a volume of up to $120\,\mathrm{dB}$ (at $1\,\mathrm{m}$) [5, pp.81], which are commonly used to communicate over more than $1\,\mathrm{km}$ [5, pp.35]. Whistled languages are also widely used in circumstances of background noise, where the simple, high-frequency sound of whistling is far easier to hear than speech or shouting.

3. Related work

As of the time of writing, no research into ASR of whistled languages has been published. However, whistling has been recognised as a possible way of communicating with computers, and several simple methods of communication based on whistling have been developed.

For example, a model has been made to accurately, in real-time, detect the presence of human whistling [9]. It was used to switch the lights on and off in a room, and it functioned with such background noises as human babble, music, and car noise.

Another notable example is the MiReLa language, in which a group of words were each assigned to a sequence of musical notes, and a simple grammar to connect these words was created [10]. This language was tested by a guide robot, also named MiReLa, employed at the San Sebastián Technology Park; it used the MiReLa language to communicate with elevators, and could be given commands by a human whistling in the same language. This method of communication has proven to be effective in overcoming the noise of large crowds, as it used a higher frequency range than the one most used by human speech; since most whistled languages use a similar frequency range to the MiReLa language [5, pp.80-81], this advantage should transfer to any true AWSR model.

Based on this research, it can be deduced that the performance of AWSR will be less impacted by noisy environments than the performance of spoken speech ASR, especially when the source of the noise is human speech. Whistled languages are often used for communication in such environments, and the benefits they provide seem likely to transfer to ASR models.

4. Experimental Setup

4.1. Data Used

For this project, 62 minutes of Silbo Gomero recordings were obtained, containing 3400 words, whistled by 10 whistlers. Those recordings were cut into 529 utterances. Recordings from 4 whistlers, which together comprised 90% of all words recorded, were used as the training data, and recordings from 6 whistlers, which together comprised 10% of all words recorded, as the development/testing data. The development/testing data set is smaller than recommended, but all other whistlers had a share of 10% or more in the data set, and would dominate the testing data if included.

All of the Silbo Gomero recordings were provided by Francisco Javier Correa, Coordinator of the Silbo Gomero Teaching Project (*Proyecto de Enseñanza de Silbo Gomero*), who works for the Ministry of Education of the Canary Islands. They were produced for educational purposes, and therefore can be assumed to represent expert use of the language. Francisco Javier Correa approved this use of the data, and informed me about the terms under which it could be published.

All of the recordings from the training data set were published under a CC BY-NC 4.0 licence; as of the time of writing, this is the first publicly available corpus of Silbo Gomero, or any other whistled language. It can be found at `http://www.openslr.org/137/`. Unfortunately, the development/testing data set could not be published.

To properly assess the performance of the ASR model trained on this data, a second ASR model was created, based on a comparable data set of spoken Spanish. To create such a data set, the data from the Mozilla Common Voice's Spanish Common Voice Corpus 4[1] was used. From that data set, speakers who were men (whistled languages have no significant differences in how they sound between sexes [11, pp.90-91], so single-sex data was used to simulate that in a spoken language), and who spoke with the same accent (the centro-surpeninsular accent) were selected. From that pool, a speaker for each whistler in the Silbo Gomero data set was chosen. For each speaker a subset of their utterances was selected so that

[1] `https://commonvoice.mozilla.org/`

every Spanish speaker had a similar number of words uttered as their Silbo counterpart, and the distribution of utterance lengths for every speaker was also similar. The Spanish speakers were separated into training and development/testing groups based on the group their Silbo "twins" belonged to.

The spoken Spanish data set that was created is a very exact mirror of the provided Silbo Gomero data set. This was to ensure that the most significant difference between those data sets is the type of speech—whistling vs speaking—and other differences are minimized. Unfortunately, the Silbo Gomero recordings were made by teachers for educational purposes, and, as such, are slower, more precise, and more deliberate than the more natural speech of the Spanish recordings. This difference will have to be remembered going forward.

Significantly, while the Silbo Gomero data set contains 62 minutes of recordings, the Spanish data set is around two times smaller, even though a very similar number of words is uttered in both. This can be explained by the fact that whistled languages are generally slower than corresponding spoken languages, and have clearer separation of adjacent words [5, pp.82], as well as the aforementioned educational nature of the recordings. This means that, when comparing results of the Silbo Gomero ASR model described in this paper with models trained on spoken speech, it should be compared with models trained on around 30 minutes of speech.

4.2. The experiment conducted

For each of the two data sets described above, a Word Recognition model was created, trained to recognise the words being spoken in a recording. Both of the models were trained with a very simple, 4-stage Kaldi recipe[2], which was developed by taking one of the recipes included with Kaldi and removing parts of it until the essentials shared by almost all recipes were left. The four stages were monophone training (mono), triphone training (tri1), LDA+MLLT training (tri2b) and SAT training (tri3b). Such a simple recipe was used because the goal was determining whether creating an AWSR model was possible with commonly used techniques, not optimising its performance.

4.3. Settings used in the Kaldi toolkit

Kaldi separates an ASR model into four phases: Feature Extraction, Acoustic Modelling, Lexical Modelling, and Language Modelling.

In Feature Extraction, the sound data is processed into information usable by the model; that is achieved by separating the recording into short windows, and extracting the Mel Frequency Cepstral Coefficients (MFCCs), which contain information about the sound frequencies present in those windows.

Then, Acoustic Model recognises which phonemes are likely to be represented by these MFCCs, based on each frame and several frames surrounding it; after that, Lexical Model searches for words these phonemes could create. Those models are trained with a pronunciation dictionary, which contains every word present in the language separated into phonemes, and a transcription of the data. Kaldi learns by separating words from transcriptions of training data into phonemes, and then learning how these phonemes sound based on the training data.

The Language Model chooses the most likely of the possible words, based on the previously recognised words and a corpus of the written language.

[2] https://github.com/agjak/silbo-gomero-asr/blob/main/run.sh

4.3.1. Feature Extraction

To extract data from the recordings into MFCCs, the settings shown in the Table 2 were used.

Table 2: *Settings used in the Feature Extraction phase of the Kaldi recipe in each of the two models.*

	Silbo Gomero	**Spoken Spanish**
Frequency Range	850–4200 Hz	20–8000 Hz
Cepstral coefficients	10	13
Additional pitch data	yes	yes
MFCC window width	120 ms	25 ms
MFCC window step	40 ms	10 ms

The Silbo Gomero settings shown are the ones found to give the best results, out of the ones tested. The spoken Spanish settings are the default MFCC extraction settings used by Kaldi, except for the fact that pitch data was included, so that the same recipe could be used to train both models.

Remarkably, even though in Silbo Gomero the speech is around two times slower, best results were achieved when MFCC window width and step were increased around four times; these parameters are very rarely modified when creating ASR models, regardless of the language, but modifying them was necessary to achieve a well-performing AWSR model. This may be because MFCCs are meant to represent several fundamental frequencies being present at the same time, as that is how the spoken speech functions. In whistled languages, meanwhile, only one fundamental frequency is present at any time, and information is mostly encoded in the changes of that frequency. When Kaldi determines the phoneme present in a window, it considers several adjacent windows, and when those windows encompass a wider time frame, it allows for the easier detection of changes in frequencies. It is possible that methods of sound representation other than MFCCs would result in better performing AWSR models. In this paper, however, the goal was to modify the standard procedure of creating an ASR model as little as possible, and determine whether an AWSR model could be created that way.

Other parameters modified were more straightforward; since Silbo Gomero occupies a narrow band of frequencies (around 1000–4000 Hz [5, pp.80-81]), we only need to analyze sounds from that frequency band, and we need less cepstral coefficients since the band is smaller. Information about pitch, which is used by Kaldi for ASR of tonal languages, was included, but excluding it lowers the results only slightly.

All of the parameters not mentioned in Table 2 were set at default Kaldi values for both models.

4.3.2. Acoustic Modelling and Lexical Modelling

To train the acoustic and lexical models for Word Recognition of spoken Spanish, an unedited Spanish pronunciation dictionary published by Open Speech and Language Resources[3] was used. As mentioned in subsection 2.2, many Spanish consonant sounds are indistinguishable when whistled; because of that, to train the acoustic and lexical models for Word Recognition of Silbo Gomero, this dictionary was modified, so that the phonemes that sounded the same in Silbo Gomero were represented by the same character in the dictionary. The transcriptions of the recordings were unedited from what was provided in the data.

[3] https://www.openslr.org/34/

The underlying parameters used by Kaldi to train the HMM-GMM model were not modified.

4.3.3. Language Modelling

The language models for both Word Recognition models were the same, as they are both forms of Spanish; Kaldi was simply provided with a portion of the Wikicorpus, a Spanish text corpus[4] and created the language models with the default tools it has for that purpose. This did not require any specific modification to be done for a whistled language.

5. Results and Discussion

As shown on Figures 1 and 2, the model trained to recognise Silbo Gomero achieved higher Word Error Rate (WER) and Character Error Rate (CER) than the one trained to recognise spoken Spanish. However, it did achieve a level of speech recognition, and its performance improved with each stage of the Kaldi recipe, which means that Automatic Whistled Speech Recognition models can be created using the HMM-GMM approach to ASR.

The performance of a Silbo Gomero recognition model trained without the changes to the window width and step in the Feature Extraction stage is far worse than the performance of this model, as shown in Table 3; this implies that those changes are crucial for a functional AWSR model.

A good study to compare these results to is the one conducted by Tyers and Meyer, who analysed ASR models trained on small data sets [12]. The results achieved by the Silbo Gomero model described here are similar to those achieved by their models trained to recognise Irish and Odia, as can be seen in Table 4. However, most of their models trained on similarly sized data sets achieved better results. That being said, they utilised transfer learning in the Coqui toolkit[5], which uses Deep Neural Networks, so the results are not directly comparable.

These results show that automatic recognition of whistled speech is not only possible, but also achievable with methods used for ASR today. It's true that the Silbo Gomero model performed worse than the spoken Spanish one, and a model trained and tested on a data set of more natural whistled speech would likely perform even worse. However, it's likely that by modifying the methods used to train the ASR model further—namely, using a different system of sound representation than MFCCs—one could create a better–performing model and compensate for those differences.

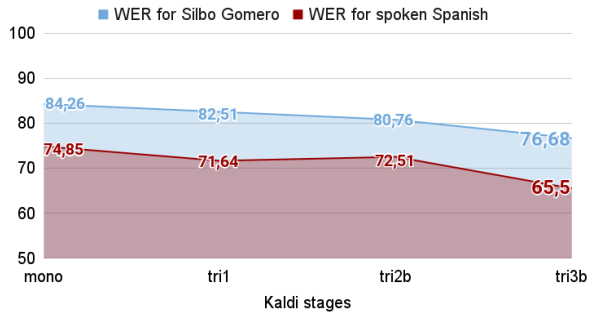

Figure 1: *WER results for the two models after the different stages of the Kaldi recipe.*

[4] https://www.cs.upc.edu/~nlp/wikicorpus/
[5] https://github.com/coqui-ai/STT

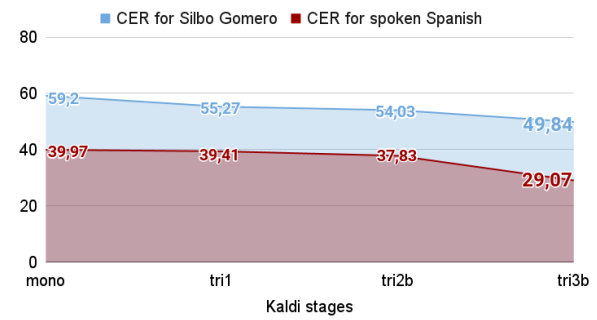

Figure 2: *CER results for the two models after the different stages of the Kaldi recipe.*

Table 3: *WER and CER results achieved by a Silbo Gomero ASR model created without the changes to the window width and step in the Feature Extraction stage.*

	mono	tri1	tri2b	tri3b
WER	90,09	110,79	119,24	117,49
CER	62,74	73,67	80,35	78,45

6. Conclusions

In this paper I showed that, when using the classic HMM-GMM method of creating ASR models, automatic recognition of whistled speech can be achieved with only minimal changes done to the training parameters. This conclusion is based on a model trained for ASR of Silbo Gomero, but the methods used should work just as well with other non-tonal whistled languages, as they are based on the same mechanism of converting speech into whistling.

This result is remarkable, especially if it could be replicated with newer, DNN-based ASR methods, and larger data sets. Developing models for accurate recognition of a whistled language would likely only require gathering a large data set of recordings, since the existing methods of accurate speech recognition could likely be adapted to accommodate whistled languages.

7. Acknowledgements

I would like to thank Francisco Javier Correa, Coordinator of the Silbo Gomero Teaching Project (*Proyecto de Enseñanza de Silbo Gomero*), for providing me with recordings of Silbo Gomero and explaining the details of how this language is used. This project would not be possible without his input.

Table 4: *Data set size, WER values and CER values for the models described and for selected models from Tyers and Meyer's study.*

Language	**Training data set size**	**WER**	**CER**
Irish	31 min, 24 s	70.73	40.57
Finnish	32 min, 29 s	60.54	30.69
Odia	32 min, 56 s	74.58	35.00
Hakha Chin	38 min, 14 s	53.28	26.48
Vallader	58 min, 58 s	54.28	26.22
Spoken Spanish	25 min, 35 s	**65.5**	**29.07**
Silbo Gomero	49 min, 25 s	**76.68**	**49.84**

8. References

[1] Grozdić and S. Jovičić, "Whispered speech recognition using Deep Denoising Autoencoder and inverse filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2313–2322, 2017.

[2] J. Pohjalainen, T. Raitio, S. Yrttiaho, and P. Alku, "Detection of shouted speech in noise: human and machine," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, p. 2377—2389, April 2013. [Online]. Available: https://doi.org/10.1121/1.4794394

[3] S. Basak, S. Agarwal, S. Ganapathy, and N. Takahashi, "End-to-end lyrics recognition with Voice to Singing Style Transfer," 2021. [Online]. Available: https://arxiv.org/abs/2102.08575

[4] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[5] J. Meyer, *Whistled Languages: A Worldwide Inquiry on Human Whistled Speech*. Berlin: Springer, 02 2015.

[6] T. Stern, "Drum and whistle "languages": An analysis of speech surrogates." *American Anthropologist*, vol. 59, no. 3, pp. 487–506, 1957.

[7] A. Classe, "Phonetics of the Silbo Gomero," *Archivum linguisticum*, vol. 9, pp. 44–61, 01 1956.

[8] R. Hammond, "The sounds of Spanish: Analysis and application (with special reference to American English)," *Hispania*, vol. 88, 01 2001.

[9] M. Nilsson, J. Bartunek, J. Nordberg, and I. Claesson, "Human whistle detection and frequency estimation," *Image and Signal Processing, Congress on*, vol. 5, pp. 737–741, 05 2008.

[10] U. Esnaola and T. Smithers, "Whistling to machines," in *Ambient Intelligence in Everyday Life*, 08 2006, pp. 198–226.

[11] R.-G. Busnel and A. Classe, *Whistled Languages*. Berlin: Springer-Verlag, 01 1976.

[12] F. Tyers and J. Meyer, "What shall we do with an hour of data? Speech recognition for the un- and under-served languages of Common Voice," 2021. [Online]. Available: https://arxiv.org/abs/2105.04674