A Guide to Building a Diphone Speech Synthesis System for Kalaallisut*

Sophie M. Rehrig

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Arts in Linguistics

Bryn Mawr College
December 2016

Abstract

Diphone speech synthesis is the process of creating an artificial voice capable of converting text input to vocalized speech by means of concatenating diphones. While compared to unit selection, the most popular modern method of speech synthesis, diphone synthesis suffers from naturalness problems, it has the benefit of requiring fewer recorded segments and thus fewer resources and less memory. In this thesis I will present a guide to constructing a diphone speech synthesis system for Kalaallisut, the national language of Greenland. I will also discuss the importance of these systems, and the need for resources that will allow members of marginalized speech communities to construct such tools themselves.

*I would like to thank Professors Jane Chandlee and Jonathan Washington for their constructive comments and assistance throughout the process of writing this thesis. Kalina Kostyszyn and Samantha Kacir provided valuable feedback on earlier drafts, and I am grateful to them both as well.

Contents

1 Introduction **4**

2 Justification for a Making New System **5**

3 Text-to-Speech Systems **7**

3.1 Early History 7

3.2 Modern Systems 8

3.3 Diphone Synthesis 10

3.4 Commonly Used Tools 12

3.5 Relevant Previous Work 15

4 Kalaallisut **18**

4.1 Phoneme Inventory 19

4.2 Phonotactics 19

4.3 Unique Phonological Problems 20

5 Creating the voice **21**

5.1 Determining diphones 22

5.2 Rules for Conversion of Orthography into Phones 23

5.3 Non-standard Words and Loanwords 24

5.4 Prosody 25

6 Conclusion **26**

A Phones **28**

B Diphone set and recording environments **29**

C G2P Rules and Syllable Weight **39**

1 Introduction

Text-to-speech systems, which take in strings of orthographic text and output auditory speech, present a solution to a broad range of problems in today's world, and are used in a variety of situations, such as assistive technology, educational toys, and automated help lines. Advances in research mean that these systems are becoming more and more natural-sounding, and even better suited for a wide variety of applications. However, much of the work on these systems is restricted to the major languages of the world, such as the languages of European colonial powers and East Asian economic powers. Thousands of languages have no synthetic voice, let alone an entire text-to-speech system, while other languages may have a system, but of poor quality. This thesis presents a guide for building a text-to-speech system for Kalaallisut, or West Greenlandic.[1] Kalaallisut has had a system built by Theiling (2013) from recorded German speech for research purposes. It therefore lacks many of the phonemes of Kalaallisut, in particular the uvularized vowels, making it a poor candidate for any practical use. In addition, it is unavailable to the general public save for a few sample outputs.

I begin in §2 by illustrating the deficit of intelligible, portable text-to-speech systems, and, in laying out my rationale for my approach, I advocate for meeting this need. In §3 I first introduce the general field of speech synthesis, briefly describing the various methods that may be used before proceeding into a more in depth discussion of the method to be used in this case, that of diphone synthesis. I then examine recent work done on languages without commercially available text-to-speech systems. §4 contains phonetic and prosodic information about Kalaallisut, including some discussion of major unique features which will present special difficulty. In §5 I present the information necessary for the actual creation of the system, such as the list of diphones, other segments included and justification for straying from the purely diphone path, rules for converting

[1]While most academic writing uses the latter name, I have elected to use the former, used by both the speech community itself and as the basis for the ISO code, throughout this thesis.

the orthography to a phonetic representation, and rules for applying prosody. In §6 I conclude the paper by presenting the steps needed to make this theoretical system into a reality, and proposing a method to increase the production of diphone synthesis systems for indigenous or threatened languages.

2 Justification for a Making New System

Crystal (2000) suggests that there are six conditions that can reverse language decline, naming as the sixth of these the ability of the language community to use modern technology. Thomason (2015) also identifies the melding of linguistic heritage and modern technology as critical, noting the increasing number of websites and applications dedicated to minority languages and increasing use of those languages on social media as key to their survival or revival. Michelle Judy Whitstone, one of the voices for the Rosetta Stone Navajo project, said in the video made about the product, "[Navajo] will begin to diminish if we don't use it anymore, and the only way, sadly, to keep it alive is through modern technology, because that's what kids are interested in doing. That's what is appealing to the younger generation, is what's on the big screen". Whitstone is expressing a variation on the popular sentiment that children are obsessed with technology, that it is essential to include the Navajo language in modern technology if Navajo children are to learn it, and to use it.

While some technological services, like the Firefox web browser, do have language options for some minority languages, others, like the Windows operating system and its associated browsers Internet Explorer and Microsoft Edge, largely include only national or majority languages. Therefore, speakers of languages unavailable through these services are required to use a majority language if they wish to navigate a computer or the Internet. In many cases, the speakers of this majority language sought to actively discourage use of the minority language not too long ago, so presenting as diverse a set of language options as possible has importance.

Combining these facts about the limitations in language availability, the logical connection between language exposure and language use, and Whitstone's comment about the increasing prevalence of technology in the modern life, one solution to the decreasing use of a heritage language by community members would be to increase the availability of that language as the primary language for an application. Making a Tongva GPS, for example, would expose community members to directional words. Making a Navajo language version of the Windows operating system would allow the speaker grandparents to teach their grandchildren Navajo, as the grandchildren teach their grandparents computer literacy.

However, one must be careful to not let the desire to increase language use and awareness cross into a paternalistic savior attitude. Thomason (2015) describes how in 2006, the indigenous Mapuche people of Chile planned to sue Microsoft for translating Windows into the (non-endangered) Mapuzugun language, without discussing it with, or even disclosing it to, the majority of the Mapuche. Microsoft responded that they were only trying to help, but to the Mapuche, it felt like Microsoft was saying they were incapable of connecting their modern lives and their language themselves. Mapuche leader Aucan Huilcan said to Long (2006), "We feel like Microsoft and the Chilean Education Ministry have overlooked us by deciding to set up a committee without our consent, our participation and without the slightest consultation." One possible way to avoid a repeat of such a situation is to make sure the language communities have the resources to implement such applications in their languages themselves, as discussed at the end of this paper. It is also important to clearly lay out the goals and desires of each party at the beginning, to make sure there are no surprise disagreements. For example, many software engineers who work on humanistic projects highly value open-source code and resources, but endangered language communities often feel a sense of ownership over their language. At all times, those of us who are working to implement speech technology systems should remain in close and open dialogue with all members of the speech community.

While Kalaallisut recently did become the national language of Greenland, until not too long

ago that position was occupied by Danish. Additionally, while many indigenous Greenlanders speak some Danish, many ethnic Danes living in Greenland cannot speak Kalaallisut.

3 Text-to-Speech Systems

Speech synthesis has many commercial applications. GPS voices, personal assistants such as Siri and Cortana, the speaking feature of Google Translate, and automated banking phone lines all make use of text-to-speech systems. Generally, a system is specialized to a single language. However, there are now some research efforts into multilingual systems and systems which are built partially off of those for other languages, to better cover the breadth of the world's languages.

3.1 Early History

Artificial speech synthesis began as a mechanical endeavor, the history of which Flanagan (1964) briefly describes. Kratzenstein in 1779 built a machine capable of producing five different vowels. From 1769 to 1791, von Kempelen worked on a machine which could produce consonants, which varied in place and manner of articulation. Riesz built a mechanism with a reed and various keys that were pressed to simulate the constriction of the human vocal tract. His machine was capable of producing some words.

After the mechanical systems came the electric ones, which are contemporary with Flanagan (1964). These depended on the use of electric energy as the source of speech sound and of circuitry and similar manipulation to modify the acoustic signal of the electric energy into an approximation of human speech. These types of systems are used today under the name of articulatory synthesis.

3.2 Modern Systems

Generally, modern TTS systems have two components, as described in Dutoit and Stylianou (2004): a natural language processing component, which creates a regular transcription containing necessary phonetic, intonational, and rhythmic information from the input; and a digital signal processing component, which converts the information from the natural language processing stage into acoustic speech. Within the natural language processing stage there are four steps according to Dutoit and Stylianou (2004). Firstly, as laid out in Jurafsky and Martin (2009) the text is initially tokenized into sentences during text normalization or preprocessing, and then the segments of the input that are non-standard words, such as numerals, abbreviations, and acronyms, are expanded into words. Secondly, in morphosyntactic analysis, as described in Dutoit and Stylianou (2004), all possible parts of speech are proposed for each word before the text is analyzed contextually, to eliminate some part of speech possibilities, in order to allow the syntactic-prosodic parser to develop a hierarchical sentence structure to aid in prosodic determination later in the process. After morphosyntactic analysis, the text undergoes a phonetization process, also called letter-to-sound or grapheme-to-phoneme, in which the orthographic text is given a phonetic representation. This can be accomplished for some languages with the aid of pronouncing dictionaries, but for all languages there are words not found in these dictionaries that too must be resolved, either by using handwritten rules or by training a machine learning algorithm, either of which can then be implemented as a finite state transducer. In addition, some languages do not have lexicons or pronouncing dictionaries available. Finally, the now-phonetized text is analyzed once more, this time for prosody, to predict prosodic structure, prosodic prominence, and tune.

For some types of synthesis (including diphone), Jurafsky and Martin (2009) note, duration and F0 (a measure of how rapidly the glottis is moving, and therefore of the pitch of the speech) are also predicted at this stage. Previously, hand-written rules were used to predict contextual variation in duration, though the current trend is towards statistical or machine learning methods. As Jurafsky

and Martin (2009) note, however, some machine learning methods use older hand-written rules to determine identifying features to be included. To predict F0 values for each segment, target points for each pitch accent and boundary tone (discussed below in §3.4.1) are first found, and then used to predict the contour for the rest of the sequence.

At this stage, with the natural language processing done, the process diverges depending on the type of synthesis in use. The main options currently, as listed by Jurafsky and Martin (2009), are formant synthesis; HMM-based synthesis; articulatory synthesis; and concatenative synthesis, which has the subcategories of unit-selection and diphone. Most commercial text-to-speech systems use concatenative synthesis. Diphone synthesis is covered more in depth in the next section. In recent years, the trend within concatenative synthesis has been towards unit-selection, as when not limited by data or memory, it is currently the method which produces speech sounding most similar to natural human speech. In this form, multiple copies of each unit are stored in a database, and the best match for each situation is chosen. Units can be at different levels of the utterance. For example, one unit-selection system might use phonemes as a unit, and another syllables. The unit that best matches the segment to be synthesized is selected by weighing factors such as environment and prosodic structure. Acoustic properties of the segment are not artificially modified.

Formant synthesis constructs artificial waveforms using mathematical models. Each waveform, representing acoustic speech, is made of a number of formants, which indicate the pitches of speech at some point in time. Values of the relevant formants are calculated based on the phone and its environment. The benefits of this method include that it does not require storage for recorded human speech segments. However, at this time it sounds unnatural and sometimes unintelligible. HMM (hidden Markov model)-based synthesis, or Statistical Parametric Synthesis, attempts to improve on formant synthesis by using hidden Markov models to determine the most likely frequency spectrum, fundamental frequency, and duration of speech.[2] More information on

[2]A hidden Markov model is a finite state machine which uses an observation state at time t to find the likelihoods of all possible results, and select the result with the greatest likelihood. For more information see Ghahramani (2001).

HMM-based synthesis can be found in Yamagishi (2006). Concatenative synthesis, unlike articulatory, formant, and HMM-based synthesis, uses segments of natural human speech strung end to end, and thus achieves a much higher naturalness, at the cost of requiring a not insignificant amount of storage.

3.3 Diphone Synthesis

Diphone synthesis, summarized by Jurafsky and Martin (2009), is a kind of concatenative synthesis in which the unit of concatenation is the diphone. A diphone is a unit composed of two adjacent partial phones: the second half of the first phone, and the first half of the second phone. The middle of a phone is the most stable across environments, with most influence from neighboring phones occurring at the boundaries of the phone under consideration, so by taking the second half of the first phone and the first half of the second phone, the effect of the neighboring phone is at least somewhat captured, and the joins will occur at the point of least variability.

Diphone synthesis only records one copy of each segment, differing from unit selection. To compensate for the lack of options resulting from having but one copy of each diphone, pitch and duration are modified to increase the intelligibility and naturalness of the synthesized speech. One common algorithm for carrying out this task is called Time-Domain Pitch-Synchronous OverLap-and-Add, or TD-PSOLA. In a pitch-synchronous algorithm, an action is taken during each pitch period, also known as an epoch. Epochs are detected either by physically monitoring the vibration of the glottis during speech recording, or by detection on recorded speech afterwards. Once there exists a corpus with labeled epochs, a windowed frame is extracted from each epoch. This frame is then used to modify the pitch and duration of its respective epoch, before all the signals are recombined. For example, to lengthen a segment, the frame will be copied. To increase pitch, frames are overlapped, and then extra frames are added to return the segment to its previous length.

Jurafsky and Martin (2009) point out two main areas where diphone synthesis suffers relative to unit-selection synthesis. Firstly, these systems have a limited grasp on coarticulation because of the fact that by definition a true diphone system will only capture effects on immediately neighboring phonemes. This means that any effects of coarticulation that are more distant than one phone will be lost. Secondly, the segments are distorted by the necessity of modifying them for length and pitch. Therefore, compared to a unit-selection system, a diphone system will sound to the listener less natural and more recognizable as an artificial voice. Such a lack of naturalness is not so much a problem in applications such as the use of artificial voices at crosswalks or on automated call lines, for which the primary aim is that the voice be intelligible. However, problems arise when the voice is being used as a voice for someone who cannot speak, or to read out personal correspondence to someone who cannot read it themselves, as the robotic nature reminds the speaker or listener that the voice is artificial.

Despite this issue with diphone synthesis, I believe it to be the best choice for situations such as are discussed in this thesis, in which speakers do not yet have any TTS system and the language is at risk of being overtaken by other languages. Unit-selection systems, in order to be able to select a segment, must by definition store multiple copies of each segment in memory. Diphone systems, therefore, are much smaller, which is necessary for using text-to-speech systems in applications on devices that both do not have much storage and do not have cheap and/or reliable signals to connect to an outside storage bank—a not unlikely scenario for speakers of many of the world's languages, who may lack high-capacity devices, regular and reliable access to the Internet, or both. In addition, in order to have multiple copies, more time must be spent recording, which is difficult to achieve in isolated communities. To obtain high-quality recordings, either a portable sound booth much be transported in at great cost, or a speaker or speakers must spend long periods of time away from the community.

Another solution to the problem of memory could be to put more research into new methods in the hopes of developing a method more natural than diphone synthesis or formant synthesis,

but less resource-intensive than unit-selection synthesis. Such a system could be valuable in the future; however, I believe there is importance to increasing the quantitative output in the present, rather than waiting for technology to progress. While Kalaallisut is not in danger of disappearing soon, many languages are critically endangered now, and working to get the resources for creating technological tools into the hands of community members is not a task that can wait on advances in technology.

In a strict interpretation of diphone synthesis, a single copy of each diphone, including diphones where one of the phones is silence, is cut from recorded utterances and stored. To create continuous speech, the diphone segments are first modified in pitch and length based on the results of the prosodic analysis, and then strung together. While generally this method works to capture most of the relevant allophonic variation, in some cases influence from phones which are not immediate neighbors has a pronounced effect. In the context of diphone synthesis, the effect of this influence is negative, as the diphone model will not capture it, lending a choppy or generally unnatural quality to the synthesized speech. Modified diphone systems may thus include a few copies of certain diphones, for which influence from beyond the boundaries of the diphone is great enough to warrant the extra space required in storage. I do not hold to a strictly diphone approach in my proposed system, and include some triphones.

3.4 Commonly Used Tools

There are various free pieces of software available to perform many of the tasks involved in building a text-to-speech system. Several examples which are used in creating diphone-based systems are described below.

3.4.1 ToBI

The ToBI (Tone and Break Indices) model, developed by Silverman et al. (1992) and described by Jurafsky and Martin (2009), represents an utterance as a sequence of intonational phrases. Each phrase ends in one of four boundary tones, and each word within a phrase has the option to be associated with one of five pitch accents.[3] The intonational phrases can furthermore be broken down into intermediate phrases, which can also receive a boundary tone. The decomposition of intonational phrases into intermediate phrases is part of ToBI's system of break indices. The strongest breaks (break index 4) are found between intonational phrases, and the next level (break index 3) is that of the intermediate phrases. Phrase index 1 is used for a standard word boundary, and phrase index 2 for a break between that of an intermediate phrase and a word boundary in strength. Figure 1 shows a neutral reading of the sentence "Marianna made the marmalade" labeled using ToBI.

ToBI was designed for use with English utterances, but has been modified and adapted for use with other languages, such as J_ToBI for Japanese and GToBI for German. It is also used in its original form when no other tool is available, as seen in the example of Davaatsagaan and Paliwal (2007), discussed in §3.5.1.

3.4.2 Festival and Festvox

Festival is free software originally developed by Black (1994) at the University of Edinburgh which provides a framework for building speech synthesis systems, and includes a few example systems, such as for English and Welsh. It is written in C++ and includes a Scheme-based command interpreter to give the user more control. This also allows users with varying levels of comfort in the

[3] A 'peak accent,' represented by H*, is in the middle to upper part of the speaker's pitch range. The lower part of the pitch range is covered by a 'low accent,' L*. A 'scooped accent,' represented by L*+H, is a low tone on the accented syllable immediately followed by a sharp rise in pitch. A high peak target on the accented syllable preceded by a sharp rise in pitch is called a 'rising peak accent,' represented by L+H*. The final pitch accent is represented by H+!H and indicates a step down onto the accented syllable from a high pitch, used only when the preceding material is clearly high-pitched and unaccented.

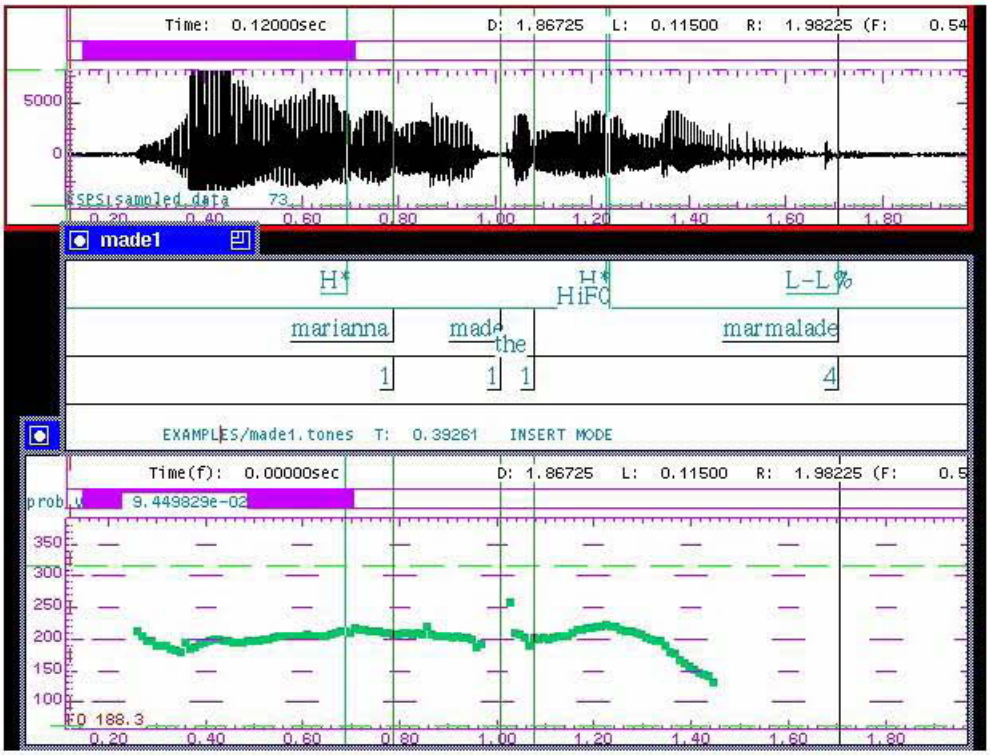


Figure 1: An example of a spectrogram labeled with ToBI conventions, from Beckman and Ayers (1994).

technology to utilize the software.

Festvox, developed by Black and Lenzo (2000) at Carnegie Mellon University, is intended to streamline the process of creating a new voice in Festival, and thus open up the field to those who might not otherwise contribute. The project also aims to increase the available documentation of artificial voices. The goal is that anyone who wishes to build a voice for any language will be able to do so.

3.4.3 MRBOLA

MRBOLA, a project run by the TCTS Lab at the Faculté Polytechnique de Mons, aims to collect and provide for non-commercial use diphone sets for as many languages as possible. The stated

goal is to provide materials for academic research on speech synthesis, especially on prosody generation. MRBOLA uses a variation on TD-PSOLA called MBR-PSOLA (Multi-Band Re-synthesis Pitch-Synchronous OverLap-Add), which is described in Dutoit (1994). It is not a text to speech system, as it is not capable of taking in raw text, but it is able to take in phonetic and prosodic information and output synthesized speech.

3.5 Relevant Previous Work

Some previous work has been done on creating diphone synthesis systems for languages lacking mainstream, openly-available text-to-speech technologies. Here, I describe a few, and also mention projects aimed at increasing access to such systems.

3.5.1 Mongolian

Davaatsagaan and Paliwal (2007) created a diphone system for the Halh (or Khalkha) dialect of Mongolian. Halh Mongolian is the national language of Mongolia, with about three million speakers in total. Some researchers before this group had explored parts of the problem of creating a text-to-speech system for Mongolian, but none had done all the work required to actually build one, making theirs the first one.

Particular challenges with which they had to contend included homograph disambiguation and phone duration. The first could be partially resolved by part of speech tagging, and the second by using a machine learning algorithm on a dataset of spoken Mongolian to train a prediction tool. As Kalaallisut seems to have no homographs, the first will not be an issue for me. They were not clear on how exactly phone duration presented a challenge. They overcame it by using statistical prediction models. Prosody in Mongolian can be predicted with good results from punctuation. They used the CART (classification and regression trees) model, a type of machine learning method

developed by Breiman et al. (1984) for creating predictions from data, to predict prosodic phrase structure and mark ToBI accents and an F0 contour. However, as there is no version of ToBI specifically for Mongolian, the team used the English model, which could yield incorrect results.

They created the system using Festival and tested it for intelligibility and naturalness on native speakers. The system performed well on measures of intelligibility in most cases, but was not considered particularly natural. Neither, however, was it particularly unnatural.

3.5.2 Arabic

Unlike Davaatsagaan and Paliwal (2007), Assaf (2005) had existing synthetic voices in her language to work with. However, she had the special challenge of working with a system in which most vowels are not written. There are existing prediction tools for filling in the missing vowels, but none is accurate enough for speech synthesis. In the end, for her system, she used a romanized transcription system of her own design, and sidestepped the vowel issue by simply including vowels in her input. However, if the public is to be able to use this system, there will need to be a different solution. Finding a standard form for input is difficult, as writing in the Arabic alphabet almost never includes the vowels, so requiring them would be a burden on the user, and there is no widely used standard romanization of Arabic. Another difficulty she encountered was with the gender of words. Some speakers of languages such as Arabic, in which some utterances are marked for if the speaker is a man or a woman, find a mismatch between the perceived gender of the synthetic voice and the gender marked on the text to be jarring.

She tested the voice she made on nine native speakers, and evaluated the system to be 85% intelligible and 59% natural. Furthermore, 78% of listeners thought the speed was correct, but also 78% found the voice at least a little annoying to listen to. The result about annoyingness is especially important to consider for any system meant to be used as part of a language revitalization program, as learners will be unlikely to make use of technology which they find unpleasant to listen

to, or difficult to bear for extended periods of time.

3.5.3 Somali

Unlike the Mongolian and Arabic examples above, Somers et al. (2006) do not create a whole diphone system for Somali. Instead, as Theiling (2013) does, they "fake" a Somali text to speech system by using German language diphones. They substituted the closest German equivalents for the phonemes missing in Somali, losing some phonemic contrasts in the process. They also ran into difficulties as Somali is a pitch accent language, with contrastive tone, such that the words for 'boy' and 'girl' vary only by whether the stress is on the first vowel or the second. They chose not to take this into account in the voice, and rather to trust context to provide any lost meaning.

For their purposes, the system was acceptable. Their primary goal was to create a system that could be used by immigrants who had not yet learned their new country's language or languages when they visit the doctor, and in this context were moderately successful, with 72% of their test group who were not wearing headphones able to understand after three repetitions, and 91% of those wearing headphones after the same number. In the context of a doctor's office visit, it is feasible to listen with headphones in a relatively quiet space, so the results for the conditions with headphones represent what the system is capable of. However, as this system is meant to be used alongside machine translations, one wonders how the errors introduced by a machine translator would affect the comprehensibility of what is already a very context dependent system. For testing, in contrast, they used grammatical Somali.

4 Kalaallisut

Kalaallisut is an Eskimo-Aleut language spoken on the western coast of the island of Greenland, with, according to Lewis et al. (2016), about 43,000 speakers in Greenland, and 7,000 in Denmark. It is by far the largest of the three major dialects of Greenlandic, and is the national language of Greenland, as well as the language of primary education there. It is on the eastern part of the language continuum of the Inuit languages, which are also present in Canada and Alaska.

While Kalaallisut is not a minority language, as a national language, it was nevertheless chosen for this project. It is one of the most widely spoken indigenous languages of North America today, and one of the few to have a standard form and standardized orthography. Additionally, as it is taught in schools in Greenland, it is more likely that speakers are able to write in the standardized orthography than with a language such as Navajo, which possesses such an orthography, but few speakers who have the ability to write it. Kalaallisut, in contrast, has a population highly literate in the standard orthography. This saves this project from the problem of creating a grapheme-to-phoneme conversion able to deal with personal transcription systems.

The large number of speakers for a single dialect of an indigenous language in North America is also unusual. I wanted to work with a North American language, as a resident of this continent, but I did not feel it was my place to be making decisions among dialectal differences, and thus imposing my view of what a standard form should be upon a language to which I have no connection, and therefore choosing a very specific variety of any other indigenous language of North America would have difficult implications. In addition to the scarcity of materials about small dialects, it is also not my place to choose one variety of the language as the one that receives new resources and attention, as this could very well also become an attempt to establish a standard. This restricted me to a few languages. Kalaallisut was the one I felt most comfortable working with ex-situ, as it is a national language of an internationally recognized sovereign nation, unlike any indigenous

language of the United States.

4.1 Phoneme Inventory

According to Fortescue (1984), Kalaallisut has 13 or 14 distinctive consonants and three vowels, three or four additional marginal consonants, and two additional optionally inserted glides. For the purposes of this system, I have chosen to leave out the optional glides. The consonant inventory includes fricatives, nasals, voiceless plosives, a lateral approximant, and a glide. Adjacent uvulars cause extensive variation in these phonemes. For a table of possible phones in Kalaallisut, see Appendix A.

Mase and Rischel (1971) defines the most important allophonic variations in Kalaallisut as aperture, length, and voice. For the first, they note that all vowels are lowered and retracted before uvulars, and /a/ is raised before other consonants in the same syllable. Length or both vowels and most consonants is contrastive. The length of a single consonant is either short or long, and within consonant clusters, in the cluster /ts/ the first consonant is long, and in any other cluster the second consonant is long. Voice is determined by manner of articulation and length. Stops and sibilants are unvoiced and nasals are voiced. Other manners of articulation are voiced when they are short and voiceless when they are lengthened.

4.2 Phonotactics

Kalaallisut has a (C)V(V)(C) syllable structure. In his discussion of phonotactics, Fortescue (1984) says that word finally, only singleton plosives and vowels are allowed, and word initially, plosives, /s/, /h/, /m/, and /n/ and vowels are allowed. Loanwords also allow /f/, /v/, /j/, /l/, and /ʁ/ word initially, in addition to permitting consonant clusters such as /kʁ/, as in *kristumiu* 'Christian."

Word medially, the cluster pattern /VʁC/ is allowed, while some other clusters occur in loanwords, as seen in the previous example with /st/.

While in general, Fortescue (1984) is agreed to have the correct analysis of which segments are allowable where, Rischel (1974) makes note of a kind of word-final reduction which increases the set of syllables which can be found word finally, which he describes as a "clipping" of the coda of the final syllable. For example, *qujanaq* 'thank you' is shortened to *qujan* with a word final /n/, usually not permissible. Rischel (1974) says that his consultants did not consider it a separate wordform, a statement backed up by the fact that the prosodic contour for the word is that of the original word with the last prosodic segment also clipped, rather than the a contour consistent with the general rule of counting backwards from the final syllable to apply tone.

4.3 Unique Phonological Problems

Phonologically, Kalaallisut is both generally simple and predictable, at least for the purpose of creating a speech synthesis system, and well-studied. One major feature that must be contended with is the sequence /VʁC/, such as in the word *erinarsuut* 'song." The uvular /ʁ/ affects both the preceding vowel, which becomes a uvular vowel, and the following consonant, which lengthens. However, as Mase and Rischel (1971) describe, there is still discussion over how much the following consonant lengthens, and whether or not the uvular is at all retained. They come to the conclusion that the uvular is partially assimilated to the following consonant, but is retained enough to cause the preceding vowel to uvularize. The duration measurements that this study took showed that the non-uvular segment is equal in length to a simple geminate consonant.

Prosodically, Kalaallisut is not well described. Arnhold (2014) says most non-final words have a high-low-high (fifty percent), high-low (forty-five percent), or low-high (three percent) contour, with the remaining two percent having no tonal contour. Speakers were found to be

more likely to use high-low-high contours when speaking slowly or reading a text, as opposed to having a dialogue, suggesting that high-low-high is the underlying form, and other realizations are reductions of that. Final words in yes-no questions having a falling pitch, while all other final words have an underlying high-low-high structure, in which the final high is sometimes reduced to a low. However, what is meant by this tonal contour is under debate, as is it agreed that Kalaallisut has no stress or lexical tone.

Overall, the phonology of Kalallisut, with the except of the /VʁC/ segments, should not pose a problem for the creation of a diphone speech synthesis system. The phoneme inventory is not terribly large. While the inclusion of the "clipping" phenomenon could make the speech sound more realistic, it is not strictly necessary, as it is primarily found in spoken language, and text-to-speech systems, by definition, have a written base. The prosody, however, could provide difficulties.

5 Creating the voice

To create the synthesized Kalaallisut voice, the possible diphones must be first determined. Then, rules for converting the orthographic representation into a phonetic representation must be created. Non-standard words have to be expanded and converted, and loanwords which do not conform to Kalaallisut phonotactics must also be converted, though not through the same set of rules. Input, in the form of words in the standard orthography, are both converted into strings of phones and broken up into syllables in order to prepare for the step of marking prosody. The prosody-marked phonetic string is then fed to the actual synthesizer.

The segments to be synthesized are prepared by making a list containing all required diphones embedded in words. Either real words or nonsense words may be used. Nonsense words, as unfamiliar words which are pronounced more carefully, have the advantage of carrying less emphasis during recording, and if for each type of diphone, each diphone is embedded in the same nonsense

word, as in Davaatsagaan and Paliwal (2007), the speaker will use a steadier rhythm. For purposes of this paper I am choosing real words rather than nonsense words, as I am not a speaker of the language. Mase and Rischel (1971) made the same decision when choosing words to record in their study on phoneme length, as they say due to the agglutinating and highly analyzable nature of the language, readers are likely to attempt to reinterpret a word into something familiar. However, unless a particular diphone is revealed to only occur with Kleinschmidt's "tone," (as discussed below in §5.4) I will choose words in which the diphone occurs in a non-accented location.

The words are then recorded by either one speaker, or one male and one female speaker if we wish to provide a choice of gender in the voice to the user, and then the diphones are extracted from the words, labeled, and stored in a database. When the system receives the phonetic-prosodic string, it pulls and modifies the diphones as necessary. Length as marked in the string causes a manipulation of diphone duration, and prosodic features are achieved through changing the value of F0.

5.1 Determining diphones

The list of diphones was created by making a list of all the allophones of Kalaallisut phonemes determined by a larger environment than just their immediate neighbor and adding that to a list of general pairs of phonemes, and then subtracted any pairs which are not allowed phonotactically.. More specificity is not needed because some variation is captured by the environment in the diphone itself. The diphones /kʁ/ and /st/, consonant clusters not normally allowed in Kalaallisut, were added to the list for their common usage in Christianity-related contexts.

An exception to a strictly diphone model was made for the pre-uvularized geminate consonants, which are orthographically represented <VrC> and phonetically consist of a uvularized vowel followed by some remnant of the uvular /ʁ/, followed by the consonant, which is lengthened to

nearly the length of a standard geminate in the language. For example, in the word *ersiut* 'vowel," the initial vowel /i/ has a uvular or pharyngeal quality, and the initial consonant /s/ is lengthened. However, there is dispute over how much of the uvular remains, and how much is absorbed onto either side, as recorded by Rischel (1974). Therefore, for these cases I will include triphones of this form in with the diphones.

When choosing the recording environments for the diphones, I used the method of determining syllable weight described in §5.4 to give preference to environments that place the diphone in a minimally weighted syllable, for consistency. I also chose instances of singleton vowels and consonants, except in cases in which some phoneme, such as /f/, only occurs in a geminate position, to try and keep consistent choices of length among all pairs, despite my method of hand-selection.

See Appendix B for the list of diphones and triphones with a corresponding natural environment for each.

5.2 Rules for Conversion of Orthography into Phones

The orthography of Kalaallisut was standardized in 1972 and 1973, and is the end of a chain of linguist-created orthographies, according to Rischel (1974). As such, the phonetic value of a grapheme can be determined from the grapheme and its immediate neighbors, with no ambiguity; the orthography is phonetic. Therefore, a list of handwritten rules for converting the orthographic representation into a phonetic representation is all that is necessary. The handwritten rules can then be turned into a finite-state transducer using available software packages, such as from Google. For an example of a method for turning orthographic Kalaallisut into a system that can be easily read off as diphones, see Appendix C. *Oqaasileriffik–The Language Secretariat of Greenland* (2016) has also developed a tool which converts orthographic Kalaallisut into IPA.

5.3 Non-standard Words and Loanwords

Non-standard words include various types of numbers, and abbreviations and acronyms. For example, in English, a text-to-speech system should be able to take in the sentence "The due date is Jan 12" and correctly determine to read "Jan" as "January" and "12" as "twelfth," rather than thinking "Jan" is the personal name and "12" is "twelve" or "one two." I do not know enough about how speaker of Kalaallisut write and pronounce numbers, abbreviations, and acronyms, to say anything about how these will be resolved. One method for resolving this lack of knowledge would be to ask native speakers to write short paragraphs in both informal and formal settings, and then read their work out loud. Recordings of their pronunciation of each non-standard word can be compared with what they wrote. They can also be asked to read aloud pre-existing texts or phrases commonly needed by text-to-speech systems.If a great enough number of speakers were to participate, I believe a large range of non-standard words could be captured.

At the beginning of this century Kalaallisut was still a co-official language of Greenland with Danish, and at the time of writing Greenland is still legally part of Denmark, which means that many residents of Greenland will speak both languages. Several thousand speakers of Kalaallisut live in Denmark, as well. Often, therefore, names of people and places are Danish in origin, or words in Danish, English, or some other language will be mixed in. These must be able to be accounted for, especially the Danish names and words. The phonotactic constraints of Kalaallisut prohibit any consonant clusters except for geminates, /ts/ and /rC/, whereas Danish and English contain many more clusters.

Kalaallisut provides for vowel harmony in the case of epenthesis to break up loanword consonant clusters, or to correct for vowels not present in the inventory. For example, the Danish name Søren is assimilated into the phonemic structure of Kalaallisut as Suulut, in which the second vowel, neutral in Danish, takes on the qualities of the first. An example of epenthesis is found in Knud becoming Kunuut. Names which have been modified by these processes already can be han-

dled just as if they were native Kalaallisut words. Non-assimilated words are most likely Danish in origin, so one solution for managing them would be to use a Danish system for them.

5.4 Prosody

According to Fortescue (1984), stress does not play a contrastive role in Kalaallisut. Rischel (1974) notes that the study of stress in Kalaallisut is not particularly advanced, as listeners perceive much influence from the tonal contour, making their judgments about where stress falls uncertain. Kleinschmidt (1851), cited by Rischel (1974), created a set of syllable weight rules, in which each syllable is given a value from one to five. Each short vowel in a syllable has a value of two, and a syllable-final consonant has a value of one. In this analysis, long vowels are considered to be two short vowels, and are thus worth four.

Holtved (1964) cites a letter from Kleinschmidt to Bourquin which gives the following examples of words (transcribed by Rischel (1974) into modern orthography) with their syllable weights marked:

(1)

na	nu	i	lu	mut	qi	lam	miit	tuq
2	2	2	2	3	2	3	5	3

Kleinschmidt (1851) then marks "tones" on words, but the analysis of Rischel (1974) on his comparison of "tone" in Kalaallisut words with that of German words reveals these tones to be different from the modern analysis of tone contour in the language. He identifies a main "tone" or accent, with longer words also having subsidiary accents. The main accent falls on the heaviest out of the last three syllables. If these are of equal weight, it falls on the first of these. The subsidiary accents occur in a pattern of alternating accented and unaccented syllables. See Appendix C for an automated method of marking Kleinschmidt's "tones."

Rischel (1974) analyzed the "tones" as stemming from a mora-based, rather than a syllable-based, stress-like system. Arnhold (2014) supports this analysis with the example of the two words *ataata* 'father' and *ataataa* 'his/her father,' which do not have the same contour when analyzed syllabically. The first has a falling movement on the penultimate syllable, and a pitch rise on the ultimate syllable, while the second has rising movements on both the penultimate and ultimate syllables. However, by dividing the words into moras, their similarity becomes apparent, as both have a high-low-high pitch movement on the last three moras.

Scholars such as Rischel (1974) and Arnhold (2014) agree that "stress" does not exist in Kalaallisut as a relevant category. They are supported by Jacobsen (2000), who analyzed seven words in carrier sentences, read by two speakers, and found no systematic co-occurrence of duration or pitch.

6 Conclusion

In order for a speech synthesis system designed for Kalaallisut to become a reality, high quality recordings of the listed words must first be made by a native speaker, and then the information gathered in this paper must be used alongside those recordings and software, such as Festival, for assembling the voice.

While the most apparent contribution of this thesis is the addition of a set of instructions for creating a test-to-speech system for a single specific language, I am hopeful that this thesis also provides a set of instructions for creating other sets of instructions for other language. This guide for Kalaallisut could be accomplished as it was both because of the standardized nature of the language, and because of the quantity of work published describing the language. Many of the world's languages, including close relatives of Kalaallisut such as Polar Greenlandic, do not have nearly the wealth of written descriptions available to Western academics and researchers as Kalaallisut

does, making a much more collaborative approach necessary. Perhaps teams composed of speech community members and native speakers, field linguists, and computer scientists can work together to create useful, human-oriented synthetic voices and other technological tools. Teams with such a makeup already exist, so the next step is to provide them with guidelines for how to accomplish the construction of a text-to-speech system in an ecnomical and effective manner.

Previous work on text to speech systems has been done primarily by engineers and computer scientists, although the webpages of both Festival and Festvox state that they hope clear and plentiful documentation will encourage those without a technical background to pursue creating artificial voices. I hope that this work will encourage more people with the linguistic know-how to isolate and analyze relevant language-specific data to prepare guides for a wide variety of languages. I think especial benefit could come from teams of linguists, engineers, and community members working together to develop language technologies for under-resourced language communities. While most work today focuses on either practical applications for widely-spoken languages or academic research into the challenges associated with less-spoken languages, I am hopeful that a collaborative, open, and people-driven effort could rapidly expand the reach and variety of language technology, providing another set of tools with which communities can fight a decline in language use.

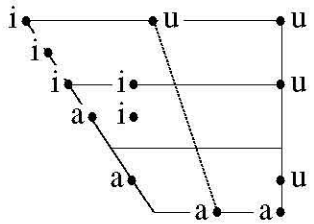
A Phones

The information in these charts is compiled from Thalbitzer (1971), Schultz-Lorentzen and Møller (1927), and Fortescue (1984) to illustrate the wide range in phonetic realization of phonemes. Variations of the same allophone are all marked with the symbol for the most commonly hypothesized underlying form.

Consonants:

	Bilabial	Labio-dental	Alveolar	Palatal	Post-Palatal	Velar	Uvular	Glottal
Plosive	p		t		k	k	q	ʔ
Palatalized Plosive			t					
Nasal	m,p		n,t			ŋ,k	q	
Fricative	v	v v	s		j	ɣ ɣ	ʁ	h
Lateral Approximate			l					
Lateral Fricative			l					
Glides	w			j				
Affricate			ts					
Aspirated Affricate			ts					

Vowels:



B Diphone set and recording environments

I have presented the phones in the below table in a phonemic way for pairs in which the phonetic realization of the phoneme is entirely dependent on the other phone in the diphone. The word given for a recording environment is written orthographically, while the phonemes are written according to the IPA. The symbol ∅ indicates a word boundary. The relevant part of each recording environment is bolded. The glosses are originally from Schultz-Lorentzen and Møller (1927), transcribed into the modern orthography and digitized by *Oqaasileriffik–The Language Secretariat of Greenland* (2016).

1st Phone	2nd Phone	Recording Environment
∅	/a/	**a**llatut 'different; in a different manner'
/a/	∅	eqi**a** 'corner of his mouth'
∅	/i/	**i**lluku 'ruin of house'
/i/	∅	ikan**i** 'over there'
∅	/u/	**u**niaaq 'that which is being trailed, dragged'
/u/	∅	pattak**u** 'piece of bone from which the marrow has been extracted'
/a/	/i/	atserp**ai** 'throw sweepings out on the dunghill'
/i/	/a/	nul**ia**q 'wife'
/a/	/u/	uu**au**araa '(the pot) is become too full of it'
/u/	/a/	ul**ua**q 'cheek'
/i/	/u/	s**iu**t 'ear'
/u/	/i/	**ui**nik 'human flesh'

1st Phone	2nd Phone	Recording Environment
/f/	/i/	allaf**fi**gaa 'writes to him'
/i/	/f/	allagars**if**figaa 'gets a letter from him'
/f/	/a/	anguf**fa**seq 'web'
/a/	/f/	anngall**af**figaa 'nods to him in the affirmative'
/f/	/u/	tiff**u**ppoq 'splashes; squirts'
/u/	/f/	ajukk**uf**figaa 'considers him/himself unworthy of it'
/u/	/ɣ/	aj**ug**aq 'superior'
/ɣ/	/i/	a**gi**aq 'violin'
/i/	/ɣ/	kiperiff**ig**aa 'longs for him with all his heart'
/ɣ/	/a/	pitu**ga**q 'made fast; harnessed'
/a/	/ɣ/	pulamm**ag**iaq 'inlet; mouth (of a fjord)'
/ɣ/	/u/	pussu**gu**ssuit 'vise'
/j/	/i/	**Ji**susip 'Jesus'
/i/	/j/	seq**ij**ak 'slack; worn-out; indifferent to his appearance'
/j/	/a/	a**ja** 'maternal aunt; aunt'
/a/	/j/	**aj**agaq 'ring-and-pin game'
/j/	/u/	a**ju**ssuseq 'wickedness'
/u/	/j/	n**uj**aq 'single hair'

1st Phone	2nd Phone	Recording Environment
∅	/k/	**k**akammik ‘surprising, too bad’
/k/	∅	kakki**k** ‘snot’
/k/	/i/	a**ki**mmippoq ‘strikes against it’
/i/	/k/	m**ik**iaq ‘fermented food’
/k/	/a/	na**ka**latsivoq ‘hangs his head in (shame or sorrow)’
/a/	/k/	n**ak**atarpoq ‘plays ninepins’
/k/	/u/	pu**ku**suk ‘back of head; nape of neck’
/u/	/k/	tors**uk**attak ‘sound (nautical term)’
/l/	/i/	appa**li**arsuk ‘little auk’
/i/	/l/	ajassaanngi**l**aq ‘be immovable’
/l/	/a/	tapita’**la**riipput ‘they are used in a lump’
/a/	/l/	tass**al**i ‘without reason; without object’
/l/	/u/	u**lu** ‘Greenlandic woman’s knife; harpoon head’
/u/	/l/	**ul**ussaq ‘iron for head of harpoon’
/ɬ/	/i/	ili**ll**ierpaa ‘puts him into a shirt’
/i/	/ɬ/	ak**ll**orippoq ‘is quick at answering’
/ɬ/	/a/	kisissaanngi**lla**t ‘they are not to be counted; countless’
/a/	/ɬ/	kuj**all**eq ‘the southern-most’
/ɬ/	/u/	i**llu** ‘peat-walled hut’
/u/	/ɬ/	k**kulloq** ‘thumb; bind-toe (of animals or birds)’

1st Phone	2nd Phone	Recording Environment
∅	/m/	**m**alaq ‘throat’
/m/	/i/	ma**mi** ‘blubber on the fleshy side of a skin’
/i/	/m/	**im**aq ‘sea’
/m/	/a/	miluu**ma**soq ‘mammal’
/a/	/m/	**am**iq ‘skin’
/m/	/u/	oqqu**mu**t ‘has found shelter’
/u/	/m/	paq**um**igaa ‘is shy of it for religious reasons’
∅	/n/	**n**ajak ‘younger sister’
/n/	/i/	**ni**giq ‘southerly wind’
/i/	/n/	palers**in**eq ‘the charred part of something’
/n/	/a/	pa**na** ‘large knife; sword’
/a/	/n/	p**an**erfaq ‘concubine’
/n/	/u/	i**nu**it ‘human beings’
/u/	/n/	n**unaat** ‘cultivated soil’
/ŋ/	/i/	saa**ngi**uppaa ‘keep in front of something with it’
/i/	/ŋ/	p**ing**eq ‘red wood (drift wood)’
/ŋ/	/a/	ua**nga** ‘I’
/a/	/ŋ/	s**ang**oorpoq ‘wriggles; winds (like a road)’
/ŋ/	/u/	ti**ngu**k ‘liver’
/u/	/ŋ/	toq**ungavoq** ‘the dead’

1st Phone	2nd Phone	Recording Environment
∅	/p/	**p**amioq ‘tail (on an animal with a round tail)’
/p/	∅	inuu**p** ‘human being (ERG)’
/p/	/i/	tussa**pi**ppoq ‘walks in darkness or on a bad road’
/i/	/p/	**ip**uppaa ‘rows it (the boat)’
/p/	/a/	**pa**lak ‘excellent; splendid’
/a/	/p/	aj**ap**erpoq ‘supports the hand against something’
/p/	/u/	**pu**juq ‘smoke’
/u/	/p/	al**up**aarpoq ‘has his wife with him on the sledge’
∅	/q/	oquppo**q** ‘is full of mold or maggots; is moldy’
/q/	∅	**q**ilaaq ‘palate; ceiling’
/q/	/i/	**qi**lak ‘sky’
/i/	/q/	al**iq** ‘harpoon line’
/q/	/a/	a**qa**jak ‘belly; stomach; abdomen’
/a/	/q/	inu**aq** ‘finger’
/q/	/u/	a**qu** ‘stern’
/u/	/q/	**uq**aq ‘tongue’
/r/	/i/	ullu**ri**aq ‘star’
/i/	/r/	millil**er**ut ‘that by which something is reduced; reef’
/r/	/a/	aappassiannngua**a** ‘my little sweetheart’
/a/	/r/	is**ar**oq ‘wing’
/r/	/u/	assalikaa**rut** ‘rolling pin’
/u/	/r/	marl**or**iarpoq ‘does something twice’
/u/	/t/	majuffapp**ut** ‘they all go up’

1st Phone	2nd Phone	Recording Environment
∅	/s/	**s**ulivoq ‘does something; is at work’
/s/	/i/	ila**si**voq ‘gets a companion’
/i/	/s/	p**i**sooq ‘rich’
/s/	/a/	i**sareq** ‘bird which has moulted’
/a/	/s/	uumma**s**ivaa ‘it came to life again in his hands’
/s/	/u/	uniu**su**ngarpaa ‘is on the point of hitting it’
/u/	/s/	tulor**us**eq ‘Frenchman’
∅	/t/	**t**iliaq ‘friend; comrade’
/t/	∅	tigutsi**t** ‘prehensile organ’
/t/	/i/	**ti**mi ‘body’
/i/	/t/	saanngut**it** ‘that with which something is strewn’
/t/	/a/	kutsisi**ta**q ‘a neglected child’
/a/	/t/	**at**ik ‘name’
/t/	/u/	**tu**nersuk ‘breast bone’
/ts/	/i/	mali**tsi** ‘follower’
/i/	/ts/	s**its**iuppaa ‘oozes through it’
/ts/	/a/	igu**tsa**k ‘bumblebee’
/a/	/ts/	mas**ats**erpaa ‘moistens; waters it’
/ts/	/u/	puku**tsu**joorpoq ‘runs stumblingly’
/u/	/ts/	ak**uts**unippoq ‘a northern bank has formed’

1st Phone	2nd Phone	Recording Environment
/v/	/i/	aana**vi**k ‘great-grandmother’
/i/	/v/	anersaam**iv**oq ‘heaves a sigh; sighs’
/v/	/a/	qaa**va** ‘his forehead’
/a/	/v/	**av**annaq ‘northerly wind’
/v/	/u/	saa**vo**q ‘is strewn’
/u/	/v/	qas**uv**oq ‘is exhausted; is tired (of waiting)’
/x/	/i/	i**ggi**ppa ‘chews the fat off of it (because it tastes good)’
/i/	/x/	ki**gg**uppaa ‘takes it with him when sinking’
/x/	/a/	ki**ggap** ‘mountain pass’
/a/	/x/	**aggaatsoq** ‘one who sleeps in his clothes; a stockfish’
/x/	/u/	i**ggut** ‘biscuit’
/u/	/x/	t**ugg**arpaa ‘cuts or hacks in it’
/h/	/i/	(no example of this sequence found)
/h/	/a/	**haa** ‘look!’
/h/	/u/	**hu**j ‘an exclamation when feeling cold’
/k/	/r/	**kr**istumiu ‘Christian’
/s/	/t/	kris**t**umiu ‘Christian’

The triphones are in the table below. All instances of /Vr/ are those found in the structure /VrC/, and thus cannot be found at a word boundary.

1st & 2nd Phones	3rd Phone	Recording Environment
/ir/	/f/	qasu**erf**issarsivoq 'has round rest'
/f/	/ir/	naaf**fer**arpaa 'gradually fills or ends it'
/ar/	/f/	oq**arf**igaa 'says something to him'
/f/	/ar/	sar**far**ippoq 'there is a strong current; it is leap tide'
/ur/	/f/	tallimars**orf**igaa 'bungles with it'
/f/	/ur/	sif**for**paa 'wrings the water out of it'
/x/	/ir/	pag**ger**poq 'is sooty (a chimney or gun barrel)'
/x/	/ar/	nutag**gar**ippoq 'is quite new'
/x/	/ur/	sig**gor**titerpai 'wrings it (the washing)
/k/	/ir/	i**ker**sissimavoq 'is ajar'
/k/	/ar/	ilik**kar**seriippoq 'is slow at learning'
/k/	/ur/	i**kor**faq 'underlayer'
/ir/	/ɬ/	qit**erl**ermipput 'they pull at hooks'
/ɬ/	/ir/	aku**ller**saat 'the midmost of them'
/ar/	/ɬ/	**arl**allit 'some'
/ɬ/	/ar/	qil**lar**lissarippoq 'is easily tarnished'
/ur/	/ɬ/	s**orl**ak 'root'
/ɬ/	/ur/	nil**lor**serpaa 'let it (food or drink) stand and get cooled'

1st & 2nd Phones	3rd Phone	Recording Environment
/ir/	/m/	tann**erm**oorpoq 'draws lots'
/m/	/ir/	ti**mer**pasippoq 'lies towards the interior'
/ar/	/m/	saner**arm**ippaa 'strikes, touches him or it with the side'
/m/	/ar/	i**mar**nersaq 'a hole in the ice'
/ur/	/m/	man**orm**ippaa 'carries it under the chin'
/m/	/ur/	kingu**mor**sorpoq 'walks backwards'
/ir/	/n/	up**ern**aaq 'spring'
/n/	/ir/	ajoqersor**ner**luppaa 'teaches him badly; teaches him evil'
/ar/	/n/	inuullu**arn**erluppoq 'leads a bad, luxurious life'
/n/	/ar/	kapi**nar**tulik 'thistle; thorn shrub'
/ur/	/n/	kill**orn**uvoq 'does not do as he is wont to do'
/n/	/ur/	na**nor**luarpoq 'holds onto something; sits quite still'
/ir/	/ŋ/	**erng**useq 'ladle; drinking cup; bottle'
/ŋ/	/ir/	ilu**nger**sorpoq 'is serious, earnest; takes pains'
/ar/	/ŋ/	**arng**ilippoq 'gasps for a breath (after a fall)'
/ŋ/	/ar/	qa**ngar**suaq 'long ago, far back in the past'
/ur/	/ŋ/	tart**orng**i 'his small waist'
/ŋ/	/ur/	un**ngor**taarpoq 'has got a wort (or worts)'

1st & 2nd Phones	3rd Phone	Recording Environment
/ir/	/p/	mallus**erp**aa 'follows after him'
/p/	/ir/	si**per**neq 'buckle'
/ar/	/p/	siveq**arp**oq 'lasts'
/p/	/ar/	ta**par**poq 'dances'
/ur/	/p/	tarners**orp**oq 'walks after death'
/p/	/ur/	a**por**poq 'strikes; runs against'
/q/	/ir/	sioq**qe**rivoq 'digs in sand; plays with sand'
/q/	/ar/	sinne**qar**tippaa 'has some of it left; has not used it all'
/q/	/ur/	**qor**suk 'green; yellowish green'
/ir/	/s/	sinn**ers**eraapput 'they take turns'
/s/	/ir/	ta**ser**aq 'pool; pond'
/ar/	/s/	uummammiuss**ars**ivoq 'takes something into his heart'
/s/	/ar/	anersaamis**ar**neri 'his sighs'
/ur/	/s/	in**ors**arpoq 'cannot keep pace; must be left behind'
/s/	/ur/	qu**sor**aq 'elegant; foppish'
/ts/	/ir/	ina**tser**iippoq 'is disobliging; is slow to obey'
/ts/	/ar/	inuu**tsar**luppoq 'is faint; famishing'
/ts/	/ur/	erni**tsor**luppoq 'has difficulty in being delivered'

C G2P Rules and Syllable Weight

```
consonants = ('f','g','h','j','k','l','m','n','p',
        'q','r','s','t','v','w')
d_consonants = ('ff','gg','kk','ll','mm','nn',
        'nng','pp','qq','rr','ss','tt','LL','xx','XX')
s_consonants = ('f','g','h','j','k','l','m','n',
        'ng','p','q','r','s','t','v','w','L','x','X')
consonants = s_consonants + d_consonants
vowels = ('a','e','i','o','u','aa','ii','uu','ai')
o_consonants = ('l','v','g')


def devoice(cons):
    """
    Takes a consonant and returns the voiceless version.
    """
    if cons == 'l':
        return 'L'
    elif cons == 'v':
        return 'f'
    elif cons == 'g':
        return 'x'
    elif cons == 'r':
        return 'X'
    else:
        print("error")
```

```
def syllables(word):
    """
    Takes a Kalallisut word and returns a list
    of its syllables.
    """
    word = g2p(word)
    cur_syl = ''
    has_vowel = False
    sylls = []
    for n in range(1,len(word)-1):
        if n == 1:
            cur_syl+=word[n]
        elif word[n] in vowels:
            if has_vowel:
                sylls.append(cur_syl)
                cur_syl = word[n]
                has_vowel = False
            else:
                cur_syl+=word[n]
                has_vowel = True
        else:
            if word[n] in d_consonants:
                cur_syl+=word[n][0]
                sylls.append(cur_syl)
                has_vowel = False
                cur_syl = word[n][1]
            elif word[n+1] in vowels:
```

```
                sylls.append(cur_syl)
                cur_syl = word[n]
                has_vowel = False
            else:
                cur_syl+=word[n]
        if word[n+1] == '#':
            sylls.append(cur_syl)
            cur_syl = word[n]
            has_vowel = False
    return syllsr_syl = word[n]
            has_vowel = False
    return sylls


def g2p(word):
    """
    Takes in a orthographic representation of a word
    and converts it to a phonemic representation.
    """
    word = word + "#"
    sound_list = ['#',]
    n=1
    while n in range(1,len(word)):
        if word[n-1] in consonants:
            if word[n-1] == 'r' and word[n] in consonants:
                if word[n] in o_consonants:
                    vc = devoice(word[n])
                    sound_list.append(word[n-1]+vc)
```

```
                sound_list.append(vc)
            else:
                sound_list.append(word[n-1]+word[n])
                sound_list.append(word[n])
            n=n+1
        elif word[n-1] == 't' and word[n] == 's':
            sound_list.append(word[n-1]+word[n])
            n=n+1
        elif word[n-1] == 'n' and word[n] == 'g':
            sound_list.append(word[n-1]+word[n])
            n=n+1
        elif word[n-1] == 'n'
           and word[n] == 'n' and word[n+1] =='g':
            sound_list.append('ng')
            sound_list.append('ng')
            n=n+2
        elif word[n] in vowels or word[n] == '#':
            sound_list.append(word[n-1])
        elif word[n] in consonants and word[n-1] == word[n]:
            vc = devoice(word[n])
            sound_list.append(vc+vc)
            n=n+1
        else:
            sound_list.append(word[n-1])
    else:
        if word[n-1] in vowels:
            if word[n-1] == word[n] or
```

```
                    (word[n-1]== 'a' and word[n] == 'i'):
                        sound_list.append(word[n-1]+word[n])
                        n=n+1
                    else:
                        sound_list.append(word[n-1])
            n=n+1
        sound_list.append('#')
        return sound_list

def print_syl(word):
    """
    Takes a word in the form of a list of syllables,
    and returns it as a string
    with syllabels separated by a period character.
    """
    sylls = syllables(word)
    retval = sylls[0]
    for n in range(1,len(sylls)):
        retval+='.'
        retval+=sylls[n]
    return retval

def weight(syl):
    """
    Takes a syllable and returns its
    weight according to Kleinschmidt and Rischel.
    """
```

```
    weight = 0
    for s in syl:
        if s in vowels:
            weight+=2
    if syl[-1] in consonants:
        weight+=1
    return weight


def weigh_syl(word):
    """
    Takes a word and returns a list
    containing syllables and their corresponding weights.
    """
    ws = []
    syls = syllables(word)
    for s in syls:
        ws.append((s, weight(s)))
    return ws
```

References

Arnhold, Anja. "Prosodic Structure and Focus Realization in West Greenlandic". In: *Prosodic Typology II: The Phonology of Intonation and Phrasing*. Oxford University Press, 2014.

Assaf, Maria Moutran. "A Prototype of an Arabic Diphone Speech Synthesizer in Festival". Sweden: Uppsala Universtet, 2005.

Beckman, Mary E. and Gayle M. Ayers. *Guidelines for ToBI Labeling*. Version 2.0. 1994.

Black, Alan. *The Festival Speech Synthesis System*. 1994.

Black, Alan and Kevin Lenzo. *Festvox*. 2000.

Breiman, Leo et al. *Classification and Regression Trees*. Wadsworth and Brooks/Cole Advanced Books & Software, 1984.

Crystal, David. *Language Death*. Cambridge University Press, 2000.

Davaatsagaan, Munkhtuya and Kuldip K. Paliwal. "Diphone-Based Concatenative Speech Synthesis System for Mongolian". In: *IMECS 2008*. Hong Kong, 2007, pp. 276–279.

Dutoit, Thierry. *A Comparison of Four Candidate Algorithms in the Context of High Quality Text-to-Speech Synthesis*. Facult'e Polytechnique de Mons, 1994.

Dutoit, Thierry and Yannis Stylianou. "Text-to-Speech Synthesis". In: *The Oxford Handbook of Computational Linguistics*. Ed. by Ruslan Mitkov. Oxford University Press, 2004. Chap. 17.

Flanagan, J. L. *Speech Analysis Synthesis and Perception*. 2nd ed. Kommunikation und Kybernetik in Einzeldarstellung. Springer-Verlag, 1964. ISBN: 3-540-05561-4.

Fortescue, Michael. *West Greenlandic*. Croom Helm Descriptive Grammars. Croom Helm, 1984.

Ghahramani, Zoubin. "An Introduction to Hidden Markov Models and Bayesian Networks". In: *International Journal of Pattern Recognition and Artificial Intelligence* 15.1 (2001), pp. 9–42.

Holtved, Erik. *Kleinschmidts Briefe an Theodor Bourquin*. Meddelelser im Grønland. 1964.

Jacobsen, B. "The Question of 'Stress' in West Greenlandic: An Acoustic Investigation of Rhythmicization, Intonation, and Syllable Weight". In: *Phonetica* 57 (2000), pp. 40–67.

Jurafsky, Daniel and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. Pearson Prentice Hall, 2009. ISBN: 978-0-13-187321-6.

Kleinschmidt, S. *Grammatik der Grønländischen Sprache*. G. Reimer, 1851.

Lewis, M. Paul, Gary F Simons, and Charles D. Fennig, eds. *Ethnologue: Languages of the World*. 2016. URL: https://www.ethnologue.com/language/kal (visited on 09/29/2016).

Long, Gideon. *Chilean Mapuches in language row with Microsoft*. Nov. 2006. URL: http://www.reuters.com/article/us-chile-mapuches-microsoft-idUSN2238412220061123 (visited on 10/24/2016).

Mase, Hideo and Jørgen Rischel. "A Study of Consonant Quality in West Greenlandic". In: *Annual Report of the Institute of Phonetics at the University of Copenhagen* (5 1971), pp. 175–247.

Oqaasileriffik–The Language Secretariat of Greenland. URL: https://oqaasileriffik.gl/en/ (visited on 11/24/2016).

Rischel, Jørgen. *Topics in West Greenlandic Phonology: Regularities Underlying the Phonetic Appearance of Wordforms in a Polysynthetic Language*. Akademisk Forlag, 1974.

Schultz-Lorentzen, Christian Wilhelm and Fru Aslaug Mikkelsen Møller. *Dictionary of the West Greenland Eskimo Language*. C. A. Reitzels, 1927.

Silverman, K. et al. "ToBI: A Standard for Labeling English Prosody". In: *Second Internatinal Conference on Spoken Language Processing*. Vol. 2. Oct. 1992, pp. 867–870.

Somers, Harold, Gareth Evans, and Zeinab Mohamed. "Developing Speech Synthesis for Under-Resourced Languages by "Faking it": An Experiment with Somali". In: *LREC 2006*. 2006, pp. 2578–2581.

Thalbitzer, William. *The Eskimo Language*. The Shorey Book Store, 1971.

Theiling, Henrik. *Kalaallisut (Greenlandic)*. Jan. 2013. URL: http://www.theiling.de/kalaallisut/ (visited on 10/15/2016).

Thomason, Sarah G. *Endangered Languages: An Introduction*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2015. ISBN: 978-0-521-68-453-8.

Whitstone, Michelle Judy. *Rosetta Stone® Navajo: Learn to Speak Navajo*. Rosetta Stone. 2011. URL: https://youtu.be/uqN3wYR-Wrg?t=3m55s.

Yamagishi, Junichi. *An Introduction to HMM-based Speech Synthesis*. 2006.