Towards
Speech Synthesis
for Icelandic

Language Technology MA Dissertation
October 2004

Björn Kristinsson
Supervisor: Prof. Eiríkur Rögnvaldsson

UNIVERSITY OF ICELAND

Towards
Speech Synthesis
for Icelandic

Language Technology MA Dissertation
October 2004

Björn Kristinsson
Supervisor: Prof. Eiríkur Rögnvaldsson

UNIVERSITY OF ICELAND

Til mömmu, pabba,
ömmu og Andra
sem gerðu endasprettinn bærilegri

Table of contents

TABLE OF CONTENTS III

ACKNOWLEDGEMENTS V

INTRODUCTION 1

1. ICELANDIC PHONETICS 3

1.1 SAMPA 3

1.2 Phoneme Length 6

1.3 Icelandic Speech Synthesis 8

2. SPEECH SYNTHESIS 10

2.1 Types of Synthesisers 10

2.2 Diphone Systems 12
2.2.1 Icelandic Diphones 13
2.2.2 Problematic Diphones 15

2.3 Concatenation of Diphones 17
2.3.1 PSOLA 17
2.3.2 Pitch and Timing 18
2.3.3 Modifications to PSOLA 18

2.4 From Text to Phonemes 19

2.5 Prosody 23
2.5.1 Naturalness in Synthesised Prosody 24
2.5.2 Prosody for Synthesis 25
2.5.2.1 Stress and Duration 26
2.5.2.2 Tones 27
2.5.2.3 Phrases 28
2.5.2.4 Speech Synthesis 31

2.6 Icelandic Synthesis 33

3. BUILDING A PHONETIC CORPUS 35

3.1 (vef)Setur hljóðan(na) 35

3.2 Towards a Spoken Language Corpus 38
3.2.1 Data Storage 38
3.2.2.1 The Nature of Sound 38

3.2.2.2 Analogue and Digital 38
3.2.2.3 Excess Frequencies 40
3.2.2.4 Storing Data for Speech Research 40

3.3 Collecting the Data 41
3.3.1 What to Record and How Much 41
3.3.2 Labelling 42

3.4 Using the Corpus 42

4. FINAL WORDS 44

WORKS CITED 45

APPENDIX 47

Icelandic Diphones 47

Acknowledgements

I’d like to thank the people at Hex Software for the food and shelter and good company. Baris Bozkurt for his help during the making of the MBROLA synthesiser. Valdís Ólafsdóttir for her linguistic insights. Bogi Ágústsson and the people at RÚV for their generous contribution towards the corpus. Mietta Lennes for providing the praat scripts and adapting them to my project. Kristján Árnason for his input regarding prosody. And my supervisor, Eiríkur Rögnvaldsson, for all his work on this project and Icelandic language technology in general.

Introduction

In the year 2003 work was started on an **automatic speech recognition system (ASR)** for Icelandic. A group called **Hjal** (e. *babble*) was formed by Hex, a software company focusing on phone-based-services, Síminn, the largest phone operator in Iceland, The University of Iceland, providing knowledge of Icelandic speech sounds and a group of students of language technology for phonetic transcription, Nýherji, a software house specialising in enterprise solutions, and Grunnur, another company interested in phone based services. This group received a government grant, which was being offered as a part of an initiative to make Icelandic more computer friendly. The project was supervised by Scansoft Software, selected because of their focus on localisation – Icelandic being the 48th language added to their selection.

The ASR was considered a great success, and is now generally available for anyone to use. Already it has been utilised as a part of various phone services created by Hex Software. But for a phone-based service understanding the caller is not enough; the system must be able to reply. For this purpose either pre-recorded messages have been used for static data, or a speech synthesiser called **Snorri**, made by Infovox, now part of Acapela group, for dynamic data. Using dynamic data, such as news or flight schedules from the Internet, could make for very popular services, but when the speech synthesis is poor, people will be reluctant to use them. The phone-system will only be as strong as its weakest link.

After the arrival of the ASR, it is Snorri the synthesiser that is the weakest link, and improving Snorri was the original target of my research. It is widely accepted that it is through prosody where the biggest advances in speech synthesis intelligibility and naturalness can potentially be made (Dutoit 1997, Holmes & Holmes 2001, Hirschberg 2002, Syrdal et al 2000 to name but a few), and so this is where I looked initially. However, Snorri belongs to Acapela group and for the general user, in this case the developer of the phone-system, it is a black box system and little or no modifications can be made externally. And so Snorri has stayed much the same.

Nevertheless, there has been some interest to make a new Icelandic speech synthesiser, and therefore I felt prosodic research might still be of some value to aid this potential future project. To this end, I created a synthesiser for the Belgian

MBROLA system[1] for practical experimentation with prosody for synthesisers. As my research went on, it became apparent that a large corpus of speech would be of immense value for further research in this area, but also for other aspects of speech synthesis, not to mention other disciplines of language technology and other linguistic studies. Such a corpus could provide all the data required to create a speech synthesiser from scratch. Therefore, work was started on such a corpus, based on recordings from the national radio (RÚV), and although small, it has given valuable insight into both prosody of Icelandic, as well as corpus making in general.

Below I will begin by describing the state of Icelandic phonetics and phonology today, as well as the history of Icelandic speech synthesis. Then I shall be looking at the options available for speech synthesis, and suggest what I believe are the most viable strategies for Icelandic speech synthesis today and in the near future. In the third chapter I will present my suggestions for building a corpus of Icelandic speech, from recording it to annotating it to putting it to use.

[1] http://tcts.fpms.ac.be/synthesis/mbrola.html

1. Icelandic Phonetics

The basis of any speech synthesiser must be the sounds of the language it is supposed to reproduce, and therefore it is appropriate to begin by looking at the speech sounds of Icelandic.

Consonants	Stops	pʰ tʰ cʰ kʰ p t c k
	Fricatives	f v θ ð s ç j x ɤ h
	Nasals	m m̥ n n̥ ɲ ɲ̊ ŋ ŋ̊
	Laterals	l l̥
	Trills	r r̥
Vowels	Close	i ɪ ɛ ʏ ø
	Far	u ɔ a

Table 1.1 The sounds of Icelandic

Traditionally, the unaspirated stops have been transcribed as [b̥ d̥ ɟ̊ g̊], but in recent years this has been changing. While the IPA standard for phonetic transcription serves its purpose well in the literature, for language technology, it is a bit hard to manipulate. Therefore, a different method is required.

1.1 SAMPA

During the making of the aforementioned ASR system a computer readable transcription standard was needed. For this, SAMPA[2] is a popular solution, but no Icelandic adaptation of the standard existed for Icelandic. Professor Eiríkur Rögnvaldsson adapted Icelandic to the SAMPA standard (Rögnvaldsson 2003), and the same transcription standard was used for making the MBROLA speech synthesiser. Even though it worked well in both projects, and has more than proven its worth, I am tempted to suggest a few changes and present those in table 2 below.

Vowels

i	*ísland*	i s t l A n t	'Iceland'
i:	*vísa*	v i: s A	'poem'
I	*kyssa*	c0 I s A	'kiss'
I:	*kisa*	c0 I: s A	'pussycat'

[2] http://www.phon.ucl.ac.uk/home/sampa/home.htm

E	*tefla*	t0 E p l A	‘play chess’
E:	*gefa*	J- E: v A	‘give’
A	*malt*	m A l0 t	‘malt’
A:	*fara*	f A: r A	‘leave’
O	*boppa*	p O h p A	‘bounce’
O:	*holur*	h O: l Y r0	‘hollow’
Y	*burt*	p Y r0 t	‘away’
Y:	*fura*	f Y: r A	‘fir tree’
9	*börn*	p 9 t n0	‘children’
9:	*fölur*	f 9: l Y r0	‘pale’
U	*rússi*	r U s I	‘Russian’
U:	*snúa*	s t n U: A	‘turn’
ey	*seint*	s ey n0 t	‘late’
ey:	*breyta*	p r ey: t0 A	‘change’
ay	*sætta*	s ay h t a	‘settle’
ay:	*læti*	l ay: t0 I	‘ruckus’
ou	*flótti*	f l ou h t I	‘escape’
ou:	*bót*	p ou: t	‘patch’
au	*sáttur*	s au h t Y r0	‘content’
au:	*látinn*	l au: t0 I n	‘deceased’
6y	*snautt*	s t n 6y h t	‘impoverished’
6y:	*auga*	6y: G a	‘eye’
yy	*hugi*	h yy j I	‘mind’
oy	*bogi*	b oy j I	‘bow’

Consonants

p	*björn*	p j 9 t n0	‘bear’
p0	*pera*	p0 E: r A	‘pear’
t	*dreki*	t r E: c0 I	‘dragon’
t0	*teppi*	t0 E h p I	‘carpet’
k	*garður*	k A r D Y r0	‘garden’
k0	*karl*	k0 A t l0	‘(old) man’
c	*gjald*	c A l t	‘fee’
c0	*kerling*	c0 E t l i N k	‘(old) woman’
C	*hjól*	C ou: l0	‘wheel’
h	*hattur*	h A h t Y r0	‘hat’
G	*saga*	s A: G A	‘story’
x	*vigta*	v I x t A	‘weigh’
n	*nef*	n E: f	‘nose’
n0	*hnútur*	n0 U: t0 Y r0	‘knot’
N	*hringur*	r0 i N k Y r0	‘ring’

N0	*þröngt*	T r 6y N0 t	‘tight’
J	*engi*	ei J c I	‘field’
J0	*pinkill*	p0 i J0 c I t l0	‘package’
m	*mark*	m A r0 k	‘goal’
m0	*pumpa*	p0 Y m0 p A	‘pump’
T	*þurs*	T Y r0 s	‘ogre’
D	*aðall*	A: D A t l0	‘nobility’
s	*siður*	s I: D Y r0	‘tradition’
v	*vagn*	v A k n0	‘wagon’
f	*fífill*	f i: v I t l0	‘dandelion’
r	*reykur*	r ey: k0 Y r0	‘smoke’
r0	*hrista*	r0 I s t A	‘shake’
l	*lög*	l 9: x	‘law’
l0	*hlaup*	l0 9y: p	‘jelly’
j	*jörð*	j 9 r T	‘earth’

Table 1.2 SAMPA for Icelandic

One of the main changes from the original SAMPA adaptation is a clearer difference between diphthongs and monophthongs. In the new version, monopthongs are mostly uppercase (with the exception of [i]), while the diphthongs are lowercase. Furthermore, no symbol used to represent a monophthong is used for making a diphthong. For instance, [A] was written as [a] in the old adaptation, and [ay] as [ai]. While more intuitive in a sense, it required regular expressions of a varying complexity to find every instance of a monophthong in the transcribed corpus, without finding the diphthongs that use the same symbol. Making a clearer distinction like this should make working with the data easier.

Another important change is the marking of unasperated and aspirated stops. Using the pairs [p] and [p0], [t] and [t0], [k] and [k0] and [c] and [c0] rather than the previous [b] and [p], [d] and [t], [g] and [k] and [J_] and [c] better reflects the modern IPA standard of Icelandic transcription. Furthermore, choosing the unvoiced symbol ‘0’ for aspiration neatly shows many features of Icelandic pronunciation; the devoicing of trills following aspirated stops for instance. Looking at the words in (1.1) we can see how the relation between the bilabial stops is shown more clearly, and also how the aspiration ‘moves’ to the [r] to devoice it.

[p0 E s t] *pest* (pest)

[p r0 E s t Y r0] *prestur* (priest)

[p E s t Y r0] *bestur* (best)

[p r E s t Y r0] *brestur* (crack) (1.1)

Even though consonant length is not marked using this standard I will continue to refer to long and short consonants meaning consonants following short vowels and consonants following long vowels respectively. This is done to save space more than anything else.

The SAMPA standard will be use for every transcription that follows.

1.2 Phoneme Length

Historically, the length of both vowels and consonants has been marked in Icelandic. In the past few years it has increasingly been suggested however that only vowel length needs be marked as the consonant length has more to do with its perceived length compared to the preceding vowel than an actual lengthening of the consonant (Pétursson 1974, for instance). This was first put to practical use during the making of the Icelandic ASR system in 2003, and did not cause any problems there. As far as the machine was concerned, there was no significant difference between what had traditionally been considered long and short variations of a consonant.

This was also done during the making of the Icelandic MBROLA synthesiser, where it worked for the most part. I say for the most part, as the synthesis of short vowels and following long consonants sounded strange in certain contexts. Generally speaking the desired effect could easily be reached by explicitly stating the length of the consonant. But for certain context, no matter how the length is modified in the MBROLA synthesiser engine, it always sounds slightly wrong. I believe this has to do with the rapidity of the transition from vowel to consonant, and somehow this had not been recorded properly during the making of the synthesiser.

Duration has not been the thoroughly researched for Icelandic for some time, and indeed, the significance of duration is not easily defined (Dutoit 1997:162). It is accepted that there is some difference in duration between different phonemes: they can be long or short. Their length is generally considered to be relative rather than absolute, but all long phonemes roughly equally long, and all short phonemes roughly equally short, with the possible exception of the short [r], which is exceptionally short (Rögnvaldsson 1989:47-48).

What needs to be determined is how well this holds and if each and every phoneme has its own inherent length, which is then modified according to context, or if they can be split into groups that share the same duration characteristics.

A brief survey of Icelandic speech sounds seems to support the necessity of making a clear distinction between long and short vowels, and the following consonants. The results are shown in table 1.3. The data for this experiment were acquired by a mostly automatic alignment of waveform and phonetic transcription of 11 minutes of speech from the Icelandic National Radio (RÚV). The text was phonetically transcribed by hand, based on the read text rather than the actual waveforms, so vowel length marks were based on expectation rather than practice, and so some discrepancies can be expected. The initial alignment of the transcription to waveform was made using MBROLIGN[3], then ported to praat[4] using a perl script.

Vowels	Median	Max	Min	Consonants	Median	Max	Min
9	70	119	47	**C**	39	100	33
9:	74	185	50	**D**	50	140	13
9y	75	114	50	**G**	50	132	29
9y:	138	176	90	**J**	50	80	30
E	60	200	21	**c**	80	140	13
E:	88	110	44	**N**	50	70	18
I	60	215	16	**T**	70	182	20
I:	81	132	64	**p**	70	150	29
O	67	160	30	**c0**	80	150	50
O:	85	106	60	**t**	60	170	22
Y	50	202	27	**f**	70	160	21
Y:	90	170	76	**k**	50	197	16
A	60	237	28	**h**	50	170	15
A:	115	144	50	**j**	45	90	13
ay	80	180	47	**k0**	92	150	39
ay:	85	90	80	**l**	50	110	16
au	83	200	50	**l0**	55	103	38
au:	139	190	50	**m**	60	170	16
ey	91	143	50	**m0**	58	66	50
ey:	90	117	69	**n**	50	212	16
i	51	172	27	**n0**	54	110	20
i:	60	66	21	**p0**	79	120	50
ou	70	184	34	**r**	50	130	16
ou:	90	215	50	**r0**	50	110	17
U	69	150	32	**t**	51	110	17
U:	91	124	50	**s**	74	177	22
				t0	70	158	28
				v	50	110	11
				x	51	70	50

Table 1.3 Length of Icelandic speech sounds.

[3] MBROLIGN is a phoneme to waveform aligner that can then synthesise the utterance using the intonation. http://tcts.fpms.ac.be/synthesis/mbrolign/mbrolign.html.
[4] praat is a free program for phonetic research. http://www.fon.hum.uva.nl/praat/

What is striking is the discrepancy in length of consonants. This shows that when looking at consonant length, context is of great importance. The longest consonants follow short vowels, and this context must be preserved for the data to make any sense. We should find that consonants after a short vowel are of roughly the same length (150-300 ms), or that the ratio between the consonant and the preceding short vowel is similar throughout (3:5 - 4:5) (Rögnvaldsson 1989:47-48). Further research in this area should be quite interesting.

The significance of duration will be discussed further in chapter 2.4.2 below.

1.3 Icelandic Speech Synthesis

The first serious attempts for a commercial speech synthesiser in Iceland were made in the year 1986 when Kjartan Guðmunsson, a computer science graduate, tried to adapt an American system to Icelandic. After six months of hard work with little results, it was clear that the American system would be unsuitable, and the project was halted for more than two years. It was decided that making an Icelandic speech synthesiser in Iceland would be an insurmountable task, and so it was decided that foreign aid would be needed. In 1989 Guðmundsson, along with linguist Pétur Helgason, started work anew, this time with the help of Swedish software company Infovox, using the Infovox system that had already been adapted to other languages. The project took just over a year, at the end of which a fully functional formant synthesiser[5] was ready. All of the above was taken from Helgason's (1990) report on the project.

By comparison, making a basic MBROLA synthesiser took about three months with no expert aid other than that gathered from various help files, although this synthesiser was unable to read anything other than phonetically and prosodically transcribed text. With the experience gained, a new MBROLA synthesis database could be created in a week or two. While there is still a long way to go, I believe this shows well how much conditions have changed. But whether they have changed enough so what was considered impossible 18 years ago is now possible remains to be seen.

The Infovox synthesiser described by Helgason (1990) was the predecessor of the current synthesiser, Snorri, who is diphone based like the MBROLA synthesiser. The method used to create Snorri is unknown to me, however.

[5] This, and other types of synthesisers, will be described in some detail in the following chapter.

It is interesting to see what Helgason (1990)[6] sees as the potential future for synthesisers. He could foresee the potential for phone-based services, but goes on to point out that as the computer cannot understand the user, this interaction will be rather one sided without the aid of a keyboard. He therefore suggests that without speech recognition, speech synthesis for the purposes of delivering dynamic data to the user will be very limited.

Now it seems that the situation is opposite to what it was: the system can understand the user quite well, but getting the data to the user so he or she can understand properly is the problem. This can only be overcome with a better synthesiser.

[6] “Nú opnast sá möguleiki að menn geti fengið upplýsingar úr gagnabanka gegnum síma. Sá böggull fylgir þó skammrifi að þótt tölvan geti talað þá getur hún ekki hlustað.”

2. Speech Synthesis

There are many ways of producing synthetic speech, and so it is important to have a basic grasp of what types of solutions are available. I will begin by looking at two of these types briefly, before focusing on one strategy in more detail, namely diphone synthesis. I shall then move on to other aspects of speech synthesis, and while diphone synthesis will remain in the foreground, much of what follows also applies to other strategies.

2.1 Types of Synthesisers

There are two main types of speech synthesisers in use today: **formant synthesisers** and **concatenative synthesisers** (Dutoit 1997:175).

Rule based formant synthesis is a complete synthesis of speech sounds, where all the sounds are generated by computer based on a large database of rules of speech production. In other words, speech is achieved "[e]xplicitly, in the form of series of rules that formally describe the influence of phonemes on one another" (Dutoit 1997:177). These systems have the advantage that as the computer generates all the aspects of the voice, changing a number of parameters can effectively create a completely new voice, even of a different gender if desired. The prosody is also fairly easy to control, as well as the speed of speech can be made much faster than that possible with concatenative systems while remaining fairly intelligible.

Even though formant synthesis has its roots in the earliest attempts at speech synthesis, and has been around for much longer than the relatively recent concatenative systems, it has so far been unable to compete with the concatenative systems' naturalness, and takes much longer to make and at a higher cost. Work is ongoing in the field, however, and recent attempts at using "higher level parameters", where the developer can control parameters that are fairly easy to understand, leaving the actual calculation of new formant positions to the computer (Carlson & Granström 1997:774), show that formant synthesis is headed in the right direction. If a breakthrough is made so that formant synthesis system can be as easily built as a concatenative system in the future, this will be a very exiting option.

Concatenative systems can be divided in two main groups: **diphone concatenation systems** and **unit selection systems**. Both are based on recordings of actual speech sounds, concatenated to make continuous speech. The system has very limited knowledge of the data it is handling, so speech is generated "[i]mplicitly, by storing

examples of phonetic transitions and coarticulations […], and using them just as they are" (Dutoit 1997:177). The difference between those two variants is what the recordings consist of. A diphone system consists of speech sounds in pairs so generally each recording is made up of the last half of one sound and the first half of the next, thus storing not the speech sounds themselves, but rather the transitions between them. In a unit selection system each recording can be longer, even a whole phrase, and variations of each with different emphasis or stress can be recorded. Where more than one instance of a unit exists, the best fitting one is selected in real-time. Unit selection systems are generally of a rather higher quality than diphone systems. This is for many reasons: for instance the joins between two recordings from the speech database are the areas where discontinuities can appear, and a unit selection system, containing larger units than a diphone system, has fewer joins and therefore fewer discontinuities. In diphone systems the density of concatenation points is much greater, or effectively one per phoneme (Dutoit 1997:187). But the higher quality of unit selection systems comes at a price as they also require many more recordings and take up much more space. This is simple mathematics. In a diphone database for a language with 30 separate phones, there are 30^2 possible arrangements of those phones into pairs of two. That would make 900 recordings. Thankfully many of the pairs never occur in speech so the final number of recordings should be much smaller than this. The Icelandic speech synthesiser described below, for instance, had 58 phones, which should mean 3364 diphones. In the end, it was made up of merely 2106 diphones.

A unit selection database would have a similar number of diphones, but many variants as well. Just adding a single additional unit to the 30 phone sample database mentioned above, say the word 'and', a recording of ever one of the 30 sounds in a position preceding and, and again after and, plus the word 'and' twice in a row, brings us up to a total of 961 possible recordings (31^2). The database grows very fast and exponentially with each and every added unit. With a large annotated corpus, it is possible to extract the required sounds automatically with very good results. This depends highly on the quality of the corpus, however.

In the end, all the required recordings for a diphone system can be recorded in an afternoon, while a unit selection system would cost much more in time and money. I would consider a diphone-based system to be best suited for Icelandic until a suitable corpus can be built.

2.2 Diphone Systems

The basic idea of diphones is that each segment contains a recording of the transition from the steady-state portion of one phoneme, to the steady-state portion of the following phoneme. However, even at the centres of phones there can still be significant variation (Holmes & Holmes 2001:71-72). This is because the rising and falling of formants occurs very gradually, and sometimes the effects of one phone can reach over to more than just its immediate neighbour. The phone may therefore never reach this ideal steady state, even at the centre. The discontinuities caused by this are one of the biggest problems facing diphone systems and clearly shows that the concatenation algorithms must be top notch.

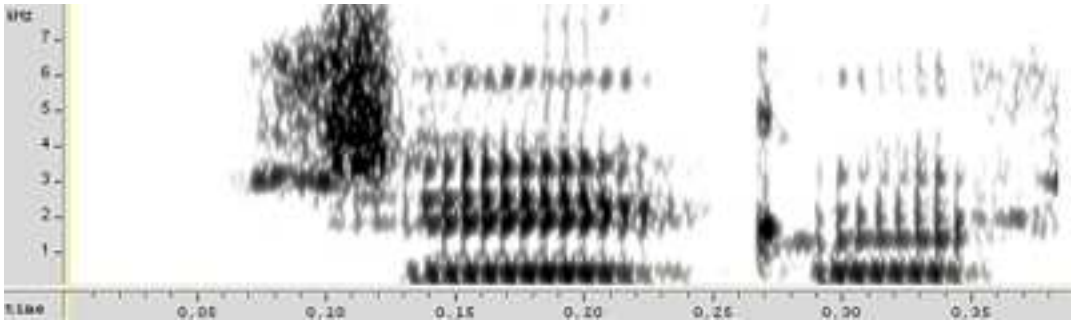

Figure 2.1 Spectrogram of the MBROLA synthesiser saying *sykur*.

(Fig. 2.1) shows a spectrogram of the synthesised word *sykur* ([s I: k0 Y r0], 'sugar'). Note how a clear line runs through the middle of the [s] (the irregular blob at the start). This is because of context differences. The [_ s] diphone was taken from the word *sögur* ('stories'), while the [s I:] diphone was taken from the word *siðir* ('traditions'). The rounded [9:] in *sögur* colours the preceding [s], while no such rounding occurs in the [s] in *hýsið*. Fortunately, mismatches of this sort are hardly audible in unvoiced fricatives such as [s]; this example was chosen as the mismatch appears more clearly on the spectrogram in [s] than many other phonemes. The same kind of mismatch can occur within every phoneme, so care should be taken to record them in as neutral an environment as possible.

Using larger units for the phones with the largest formant movements is another way to try and eliminate these discrepancies. The problem with this approach is that then we are moving towards unit selection systems, and as has already been mentioned, every added unit is exponentially costly. That said, triphones are used along with the diphones technically speaking as diphthongs are usually recorded as a half of a diphone. A balance between the two must be found.

2.2.1 Icelandic Diphones

Finding the required diphones can be tricky. First a list of a list of target diphones must be compiled. With the 58 phones of Icelandic identified (table 1.2), a list of potential pairs of phones is easily acquired. As has already been mentioned, theoretically the formula for the number of diphones required is $D = p^2$ where D is the number of diphones and p is the number of phones. In this case, $D = 58^2 = 3364$. However, many of these potential diphones can be eliminated. For instance, /J/ only ever appears before a /c/, can never be preceded by a consonant and causes diphthongisation in many vowels so only relatively few of them could ever appear before it.

A phonetically transcribed corpus would be an invaluable aid to this process of finding valid and invalid diphones in Icelandic. Today the only such corpus is the one made during the making of the Icelandic ASR, and while very helpful it only contains transcriptions of individual words. There are many more diphones that must be recorded that occur only between words or at boundaries of compounds. For instance, there are no occurrences of the diphone [9: g] in the corpus, but it can be found in context such as *sjö gallabuxur*, 'seven jeans', or in any other place where there are seven of a word beginning with [g].

There are several ways of recording the target diphones. One thing is clear, the corpus must be specialised for this purpose. Dutoit (1997:190) states that "a list of 100 phonetically balanced sentences covers only 43 percent of the 1,200 units required for French, with a redundancy of about 80 percent". It is doubtful that Icelandic is very different. There is also some debate about what type of corpus to use (Holmes & Holmes 2001:73). Is it better to use isolated words or whole sentences, nonsense words or real words? Bigorne et al (1993)[7] argue for the use of logatoms (another word for nonsense syllabic sequences), while Dutoit (1993) argues in favour of word-based units. The same applies for whether the target units should be in stressed or unstressed syllables. As Dutoit (1997:191) points out, stressed syllables are longer and thus there is less risk of coarticulation affecting the phones, decreasing the chances of discontinuities over concatenative boundaries. Unstressed syllables on the other hand are more numerous in speech, and producing them well could increase

[7] According to Dutoit (1997). He cites Bigorne et al (1993) in his text, but in his list of references, Bigorne et al are listed as being published in 1991. As the article appeared in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing 93*, I assume 1993 is the correct year.

segmental quality. The qualitative difference between vowels in stressed and unstressed syllables could also have an effect, meaning that the two might have to be treated entirely separately. The same debate applies to speaking rate, as slower speaking may result in higher intelligibility but more overarticulated units and therefore decreased naturalness.

For the Icelandic MBROLA synthesiser, a combination of mostly real words and a few nonsense words was used. The nonsense words were mostly compounds where one word ended with the first half of the required diphone, and the second word started with the second half. For instance there were made up words such as *Malmö-ást (*Malmö-love) and *biðhjól* (waiting-bike) for the diphones /9 au/ and /T C/ respectively.

I would suggest using a combination of words and nonsense words as above. As long vowels only occur in stressed syllables, then they must be extracted from those. The rest should be taken from unstressed syllables, but special care must be taken with the short vowels. The short vowels should be taken from syllables that, were they stressed, the vowel would still be short. For the diphone /O m/, for instance, 1a) in (2.1) would be less ideal than 1b). If we remove the prefix *vel-* the first syllable becomes stressed, and the /O/ in 1a) becomes /O:/ as shown in 2a). The /O/ in 1b) remains unchanged (2b)).

1a) [v E l k0 O m I n] *velkomin* (welcome, pl. neut)

1b) [v E l k0 O m n I r0] *velkomnir* (welcome, pl. masc)

2a) [k0 O: m I n] *komin* (arrived, pl. neut)

2b) [k0 O m n I r0] *komnir* (arrived, pl. masc) (2.1)

Even though actual vowel length is marked explicitly in MBROLA by the automatic phonetisation, the qualitative difference between the transitional phases of the diphones can be significant for comprehension, especially when emphasising certain syllables. If these guidelines are followed, a version of each variant should be available in the database.

A suggested list of diphones, based on the ones used to record the MBROLA synthesiser, is included in the appendix. It should be fairly extensive, but remains somewhat inaccurate until more research with the aid of a large phonetic corpus can be undertaken. There are both chances of rare diphones missing, or over-generation that would only be spotted through thorough and methodical research. A sample of this sort of over-generation that has now been found and rectified is shown in (2.1)

above, and described in 2.1.2 below, where only after making the speech synthesiser the nature of unvoicing of trills after aspirated stops became clearer. I would even go further and suggest unvoicing of other sounds as well after aspirated stops. We could, for instance, look at the word *atvinna*, 'work', which in my pronunciation could be either transcribed as [A: t0 v I n A] or, less intuitively, [A: t f I n A]. By experimenting with the MBROLA synthesiser the latter even seems to sound better, but whether this is universal or only applies in certain contexts remains to be seen.

2.2.2 Problematic Diphones

A few language specific issues must be mentioned. Firstly, most speech synthesisers for Icelandic seem to have a tendency to acquire a speech impediment called *gormæli*, which is the Icelandic equivalent of replacing [r] with [w] in English. This may be because of the nature of the Icelandic [r] combined with how the speech synthesiser produces sound lengthening. While most voiced sounds have a repetitive waveform and can therefore be lengthened by simply repeating the waveform segments, the long [r] is a trill. The mismatch between a glottal cycle on the one hand and the trill on the other distorts the sound producing something that sounds a bit like *gormæli*. This effect is hard to overcome, but hopefully the guidelines for recording short and long vowels (and therefore short and long consonants) mentioned above will suffice to minimise this problem. The short [r] is extremely short, merely a tap, and repeating that sort of waveform results in something strange, perhaps contributing to the *gormæli* effect. The combination of recording it clearly and explicitly reflecting this shortness in the synthesiser should at least somewhat decrease the number of perceived cases of this phenomenon. Glottal cycles and repeating segments are described in more detail in 2.2 below.

Syllable initial unaspirated consonants followed by /l/ or /r/ are quite troublesome: should these be transcribed as, for instance, [k0 l ou: r0] or [k l0 ou: r0] (*klór*)? Both variants were recorded for the Icelandic MBROLA synthesiser, but it turns out that the two are mostly interchangeable. I would therefore suggest that the latter way of transcribing this sort of cluster be used throughout. As described in 1.1 above, thinking of the aspiration (marked with a '0') as 'moving' over to the following liquid and devoicing it may be a helpful mnemonic during transcription.

There is also the question of dialectal variation. This should not be of any great significance for Icelandic as dialects are minimal, and two of the most common ones,

harðmæli and *linmæli* have mostly the same inventory of diphones, so sounds made from the speech of a speaker of *harðmæli* should be useable to make a speech synthesiser speak *linmæli* and vice versa.

One rather rare but useful dialectal variation on the north coast means that /N/ is always and only followed by [k] as shown in (2.2), which decreases the number of required diphones a bit. Whether this is acceptable though is arguable, and while I have this dialectal feature, I recorded the more widely used variant anyway.

-ngl- variant	*standard*	*gloss*	*translation*	
[k r0 i N **k** l A]	[k r0 i N l A]	*kringla*	‘disc’	
[U l i N k Y r0]	[U **N** l i N k Y r0]	*unglingur*	‘teenager’	(2.2)

Allowing for some of the dialects can have an advantage, however. One of those dialects is the ‘voiced variant’ of the north coast (*raddaður framburður*). It differs from the more common dialects in that post-stop nasals and /l/ are voiced. While the dialect is fairly rare, accommodating for this difference will also enable to the synthesiser to pronounce foreign words more clearly. Otherwise it would be forced to pronounce words like ‘help’ as [h E l0 p] and ‘banker’ as [p au J0 c E r0].

Which brings us neatly to another dialectal variant. Before ‘ng’ and ‘nk’ in Icelandic certain monophthongs are diphthongised. For instance, *banki* (bank) is pronounced [p au J0 c I] and *drengur* (boy) is pronounced [t r ei N k Y r0]. This explains the [p au J0 c E r0] pronunciation of ‘banker’ above. The ‘West-fjord monophthong variant’ (*vestfirskur einhljóðaframburður*) does not have this diphthongisation, and if diphones for that dialect are included as well, most foreign words can be pronounced. Pronounced with a strong Icelandic accent, no doubt, but still fairly intelligibly.

It should be noted that realistically the dialects can hardly be combined. With the West-fjord monophthong variant you can pronounce an [A] before *nk* in spelling, but then you must devoice your /n/. However, with the voiced variant, you do not have to devoice the /n/, but you must diphthongise the preceding [A]. Recording those two dialects should nevertheless suffice, as the required diphones will exist. For instance, to pronounce ‘banker’ properly (albeit with a strong Icelandic accent), [b a J c0 E r0], we can get the [a J] diphone from the monophthong variant and the[J c0] diphone from the voiced variant.

standard	*m.phth. variant*	*voiced variant*	*gloss*	*translation*	
[b au J c I n]	[b a J c I n]	[b au J c I n]	*banginn*	‘coward’	
[b au J0 c I]	[b a J0 c I]	[b au J c0 I]	*banki*	‘bank’	(2.3)

Having the option of changing dialects within Icelandic is perhaps not something that would be worth putting a lot of effort into, as the variation is so slight and little or no social stigma is attached to any of them. But if, as a side effect, more cross-language intelligibility can be acquired, it may be worth the effort to record the additional diphones required to cover those variants. Precisely because the variation is so slight, and the differences are fairly clear and covering only a few easily definable phonetic contexts, the increase in size and complexity of the database is fairly small, but the increased flexibility will be great.

2.3 Concatenation of Diphones

The actual technical side of aligning the diphones for playback is rather outside the scope of this essay, but having some idea of what the process involves can only be helpful during the recording process. I will be looking mainly at one option and some of its variations.

2.3.1 PSOLA

Splicing two waveforms smoothly together can be quite problematic. To begin to solve this issue, we must have a basic understanding of how voicing is produced. When pronouncing voiced sounds, the vocal folds are nearly touching each other and air is forced through the glottis (the slit between the vocal folds). They begin to vibrate rapidly, and the glottis rhythmically opens and closes. Each such opening and closing of the glottis is called a **glottal cycle**. The discontinuities between two segments can be minimised if the join occurs at the same position during a glottal cycle for both segments (Holmes & Holmes 2001:74). The ideal position for joining is where the amplitude is at its lowest. Therefore, the position of each glottal closure, or the pitch pulse, is marked on the waveform and those **pitch markers** can then be used to create a windowed segment for every **pitch period**. The window should be centred on the maximum amplitude point of the pitch period, smoothly tapering to either side. The size of the window is longer than a single pitch period so some overlapping will occur. After the tapered waveforms have been overlapped they can be added together, creating a new waveform.

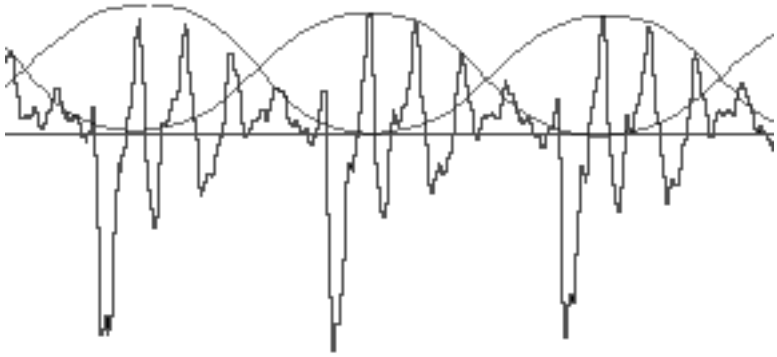

Figure 2.2 A tapered window centered on the maximum amplitude is applied to each pitch period. (Based on Holmes & Holmes 2001:75)

This process is a combination of two techniques: **pitch-synchronous** joining, and **overlap-add**, and is therefore called **pitch-synchronous overlap-add**, or PSOLA.

2.3.2 Pitch and Timing

One of the main strengths of the PSOLA method is that pitch can be modified relatively easily. It can simply be raised by decreasing the spacing between pitch markers, and lowered by increasing the spacing. This does have some effect on the synthesiser quality, however. Mostly the formant bandwidth is slightly widened, but moderate widening does not seem to be perceptually significant. If the window is wide enough, pitch change can range from half to twice the original with little detrimental effect (Holmes & Holmes 76).

Modifying timing is also relatively straightforward. To produce a longer sound unit, one simply has to repeat the pitch period for as long as is needed, using the same overlap-add process as is used to concatenate different phones.

The problem with the two methods mentioned above is that they rely on having the pitch markers in the correct positions. These can be hard to detect automatically without the aid of a laryngograph, and will always have to be corrected by hand afterwards, and it takes a trained eye to find them.

2.3.3 Modifications to PSOLA

To overcome these difficulties, various alternatives to PSOLA have been suggested and tested. One of these modifications is the **multi-band resynthesis overlap-add (MBROLA)** method, the same as was used to make the Icelandic MBROLA speech synthesiser. MBROLA is based on **multi-band resynthesis PSOLA (MBR-PSOLA)**. MBR-PSOLA works in much the same way as PSOLA, but the segments in

the database are all **resynthesised** with a constant pitch, which means every pitch period is exactly the same length. This eliminates the need for pitch markers as every segment now has the same known pitch value. This also increases intelligibility because there will be no mismatches in pitch in concatenated segments, but could also potentially decrease intelligibility if the resynthesis somehow fails. The modifications are applied only once to the entire database and therefore do not complicate or slow the synthesis in practice. MBROLA works on the same basis, but with a version of DPCM compression applied to the waveforms to make the whole system take up much less memory while retaining most of the quality.

2.4 From Text to Phonemes

With the diphones recorded and a database created what we have is a system that can read text as long as it has been phonetically transcribed. As automatic prosody generation is dependent on word classes and word context, this cannot easily be automatically imposed on transcribed text. To do prosodic generation, then, the text is examined and analysed, a prosodic model generated, and this model then imposed on the text, and the phonetic conversion automatically made afterwards.

For the purposes of my research I made a small database of strings of no more than 5 characters along with their phonetic transcription. I then used a simple perl script to read the text, starting with 5 characters and trying to find a match. If no match was found, the last character was thrown away and the program looked for a 4-character match, and so on until a match was found. Then it would try the next 5-character string until it reached the end of the text. Initially, this frequently resulted in unpronounceable sequences and so more rules were added until most of the strings returned could at least be parsed without causing program errors in the MBROLA engine. There were still plenty of pronunciation errors, however. These would have to be corrected by hand.

The quality of a system like this is hardly acceptable for anything other than basic research however, and while a better database of rules could probably be created manually, the lack of research in the relation between orthography and phonetics means that other methods might be more feasible.

Dutoit (1997) mentions two basic strategies for phonetisation: **dictionary-based** and **rule-based** strategies. The former uses an extensive phonetic dictionary to find how words should be pronounced. The phonetic dictionary created during the making of

the ASR in 2003 could very well be of some use here, although whether it is thorough enough is hard to say. The other option, rule-based strategies, is similar to the simple program mentioned above, but far more effective rules based systems can be made. A mixture of those methods is also conceivable, where most of the phonetisation process would be rule based, while a dictionary of a few hundred common words would be used when applicable. A quick survey of various texts of Icelandic showed that the 1000-2000 most common words in any text will cover roughly 90 percent of the text. It could therefore be helpful to make a large corpus of texts in an appropriate domain and transcribe the most common words by hand to aid the phonetisation.

There are two basic types of rule-based systems. First there are **expert rule-based systems**, where an expert has to build the rules by hand. As pronunciation is Icelandic is fairly regular, this option might very well be feasible. Another option is **trained rule-based systems**. While trained rule-based systems could never hope to compete with well-structured expert systems, they have the advantage that they are largely language independent and require little or no linguistic knowledge. They do require a transcribed corpus for training, however. The only transcribed corpus available is sadly lacking for the purposes of training a rule-based system as it only contains transcriptions of isolated words, not sentences and continuous speech. As mentioned above, this leaves some boundary phonemes unaccounted for. A combination of the two methods might work, however, training the system on a transcription of isolated words and creating rules that govern word boundaries by hand afterwards.

Artificial neural networks for letter to phone conversion are an exciting option used successfully for Romanian (Burileanu 2002), which was at a similar level as Icelandic speech synthesis is at today. An added bonus is that neural networks are language independent, and the importance of this will become clear later. Without the resources for practical application of this system for Icelandic, what follows will be based on the Romanian results, and speculation and expectations for Icelandic, hopefully providing a basis for practical experimentation and refinement in the not too distant future.

The input text is fed to the system, five characters at a time, with the target character at the centre and any blanks filled with a boundary character (#, in the Romanian case). The words are then shifted so the next input vector will have the next character as its target and so on until the end of the sentence is reached. Apart from simply learning what phone can correspond to each letter, the neural networks are also

trained to learn the articulatory features of each phone, enabling the system to make more abstract rules. What features are needed is hard to say without some experimentation, as Burileanu (2002:220) points out. At first these rules would be based on accepted theory for the language, but through trial and error the difficult ones causing the most errors must be removed and new 'dummy' features added that may or may not correspond to any theoretically sound features. These must be practical, first and foremost, and may or may not reflect the theory.

Another rather simpler method is using **decision trees**. Kevin Lenzo's **t2p** text-to-phoneme converter[8] can create these trees automatically from a phonetically transcribed dictionary and use it to predict the pronunciation of previously seen or unseen data. It is used by the Festival system[9], and the MBRDICO project[10], which is a text-to-speech system for MBROLA synthesisers, is based on the same algorithm. It starts by reading through the dictionary, trying to align each letter with a phoneme.

```
R E D D A Ð I S T
r E t _ a D I s t                                    (2.4)
```

(2.4) shows one possible alignment of the word *reddaðist* ('turned out well in the end'). As not every letter can be assigned a phoneme, the program will have to make a choice and align one of the letters with null, marked as _ (Lenzo 1998).

Preliminary testing seems to indicate that Lenzo's solution works quite well for Icelandic. However, adapting the program properly to Icelandic has caused some problems, and until these issues have been resolved, proper evaluation is pending. The problems are mainly two: first of all, the program reads the special characters (þ, æ, ö etc) incorrectly, and secondly, it assumes that the number of the phonemes in the phonetic strings must never be longer than the number of letters in the orthographic string. Due to phenomena such as preaspiration and inserting [t] between [s] and [n] in Icelandic, this last restriction is rather often violated. For instance, the word *epli*, 'apple', is transcribed [E h p l I], meaning there are more phonemes than graphemes, so the word will simply be ignored by the program. A way around this is to make up combined phonemes, such as [h+p], bringing the number of phonemes down to four ([E h+p l I]), and therefore fulfilling the restriction. One way around the problem with special Icelandic characters is to assign a a different more computer friendly character

[8] t2p can be downloaded from http://www-2.cs.cmu.edu/~lenzo/t2p/
[9] http://www.cstr.ed.ac.uk/projects/festival/
[10] http://www.tcts.fpms.ac.be/synthesis/mbrdico/

to each of them and replace the characters in the text with their new symbols. Initially I used numbers, 0-9, but as 0 seems to have a special function in the program, it did not work as planned. I therefore used the letter ‘Q’ as a replacement. The alignment is shown in below

```
Á Ð É Í Ó Ú Ý Þ Æ Ö
1 2 3 4 5 6 7 8 9 Q
```
(2.5)

This is not very intuitive however, but an automatic converter should be easy to implement so the user would never have to think about these adjustments.

After the initial alignment is done, **feature vectors** can be created. These show each letter in context of three preceding and three following graphemes, where L stands for the letter itself, L1 the next letter to the left, R1 the next to the right and so on (Lenzo 1998). A vector for the word *eldhúsbekkur* (or *eldh6sbekkur*, using the substitute number instead of the *ú*), meaning ‘kitchen counter’, is shown in (2.6). The underlined column shows the letter the phoneme will be aligned to.

L3	**L2**	**L1**	**L**	**R1**	**R2**	**R3**	**Phoneme**
_	_	_	e	l	d	h	**E**
_	_	e	l	d	h	6	**l**
_	e	l	d	h	6	s	**t0**
e	l	d	h	6	s	b	_
l	d	h	6	s	b	E	**U:**
d	h	6	s	b	E	k	**s**
h	6	s	b	E	k	k	**p**
6	s	b	E	k	k	u	**E**
s	b	E	k	k	u	r	**h+k**
b	E	k	k	u	r	_	_
E	k	k	u	r	_	_	**Y**
k	k	u	r	_	_	_	**r0**

(2.6)

The resulting vectors are essentially **context-sensitive rewrite rules**, where L “is the letter itself; the other ‘features’ are letters in adjacent positions” (Lenzo 1998). From this a decision tree built which should be fairly well capable of guessing what is the appropriate pronunciation of a string.

It should be possible adapting the phonetic dictionary created during the making of the ASR system, and indeed, with a few minor adjustments t2p could read the dictionary and create a decision tree that worked fairly well. Even with this rough sort

of adaptation the results look promising. With a more thorough conversion and fine-tuning, for instance finding what combined phonemes will give the best results, the results would doubtless improve. The results of a short test are shown in table 2.1. A random news article was selected from a news site, containing just over 200 words, and so we can assume that at least a few of the words could not be found in the dictionary.

Every error		*Without boundary and function word errors*	
Phonemes	Words	Phoneme	Words
94,69%	**69,72%**	**98,92%**	**93,58%**

Table 2.1 Accuracy of the t2p program for Icelandic

The results are split into two categories, both showing the number of correctly predicted phonemes on the one hand, and correctly predicted words on the other. The first category counts every error the program made, the other is more forgiving. As the dictionary is based on words being pronounced on their own and not in context many of the errors occurred on word boundaries, and in most cases these could easily be predicted with either a larger corpus, or some rules applied to the automatically derived transcription afterwards. The other common error was length in function words, such as *að* ('to'), *og* ('and') and *í* ('in'). On their own and when emphasised these could be pronounced [A: T], [O: x] and [i:], respectively. However, in normal speech the vowels are usually short, and the consonants are either dropped or can change from unvoiced to voiced, depending on the context. If we assume that these errors are easy to fix afterwards and we can therefore simply ignore them, the success rate is much higher, as shown in the second column. Results like this may be unachievable in practice, but they look promising nonetheless. Should the error rates fall somewhere in between those reported above, they must be considered acceptable by most standards.

2.5 Prosody

Speech is much more than a series of monotonous sounds stringed together to form sentences. There are things to consider such as stress, pitch, intensity, duration; even the silences have an important function. This is prosody. Prosody in speech synthesis is not only needed to make the speech more natural sounding or pleasing to the ear, but it is essential to improve intelligibility. Prosody both lends prominence to parts of an utterance through emphasis, and therefore is important to the meaning of a

sentence, as well as working as a grouping function, which makes understanding easier for the listener. In fact, it seems that the better the prosody, the more background noise can be tolerated without negatively affecting comprehension to any great degree (Fordyce 1998). That is to say, with basic and/or unnatural prosody, minimal background noise or disturbance is enough for the listener to lose track of what is being said, while with more advanced prosody it is easier to follow even under a fair amount of disturbance. If we consider that this background noise that can interrupt comprehension can be something as simple as phone line static, finding ways towards natural sounding synthesis becomes all the more important.

2.5.1 Naturalness in Synthesised Prosody

Consider the following example.

> Ég kom á bíl frá Akureyri, ekki með flugvél.
>
> I came by car from Akureyri, not by plane. (2.7)

Here it would be appropriate to put stress on the word for ‘car’, as it is contrasted by the word ‘plane’. Indeed, putting the stress on, say, ‘Akureyri’, would be unexpected and confusing. But how can the synthesiser know this? To gather this information automatically is tricky. It requires knowledge and understanding of semantics and this can hardly be expected of a speech synthesis engine at this time. This is the dilemma we face in speech synthesis. Just as prosody is important to decipher the meaning of a sentence, meaning becomes important for the proper application of prosody. As speech synthesis systems today generally have limited ways of gathering any information about the meaning of what is being said, we have to accept that full naturalness cannot be acquired now and we must do the best we can with what we have. (Monaghan 1990:89) gives a fairly reachable target to aim for:

> Acceptable intonation must be plausible, but need not be the most appropriate intonation for a particular utterance: no assumption of understanding or generation by the machine need be made. Neutral intonation does not express unusual emphasis, contrastive stress or stylistic effects: it is the default intonation which might be used for an utterance out of context. [...] This approach removes the necessity for reference to context or world knowledge while retaining ambitious linguistic goals.

However, while creating a speech synthesis system it is vital to bear in mind that while neutral intonation is acceptable for reading of previously unseen text, there are

means to improve the prosody by marking the text that the synthesiser will read. Care should be taken that while a built in default prosody for general use is required, there must be a way for the user to explicitly mark prosodic variations so that some control can be exerted over the output. The importance of this becomes clear when we contrast (2.7) above, where marking stress on 'car' explicitly would make the whole sentence easier to understand, with (2.8) below, where the meaning of the sentence can change depending on whether a phrase boundary is present or not.

1. a) *Ég ætla ekki að taka strætó | af því að ég er svo nískur.*
   'I'm not taking the bus | because I'm so cheap.'
   Meaning: It is because I am so cheap that I will not take the bus.
   b) *Ég ætla ekki að taka strætó af því að ég er svo nískur.*
   'I'm not taking the bus because I'm so cheap.'
   Meaning: The fact that I happen to be cheap is not the reason for my taking the bus. (2.8)

Unlike (2.7), (2.8) does not merely sound odd using neutral intonation, but the meaning of the utterance depends on the prosody. The neutral intonation may sound quite natural, but the meaning could be wrong. This sort of ambiguity cannot be avoided in unseen text, but if problems like these come up in a predefined text that is, for instance, part of a telephone service, the service provider must be able to communicate with the synthesiser to correct it.

2.5.2 Prosody for Synthesis

A simplified version of **ToBI (Tones and Break Indices)** will be used in the following chapters to describe Icelandic prosody. ToBI is an abstract way of describing the melodic curves of speech, using sequences of relative **tones**. The tones are defined as the "phonological abstractions for the target points obtained after broad acoustic stylisation" (Dutoit 1997:142). This abstract nature of the tones seems reflect how we perceive prosody, and could therefore be quite helpful. Indeed, there are so many ways to produce the same sentence while retaining the same emphasis, and the same general pitch contours that perhaps an abstract and general description might be more appropriate that a clearly defined mathematical formula. The actual curve of the fundamental frequency might then be generated, with perhaps a bit of randomisation thrown in, to fit the rough guidelines set by the tone values, resulting in natural

sounding speech that still sounds differently each time. In any case, using ToBI should be useful both for human readers and synthesisers.

2.5.2.1 Stress and Duration

Stress, just like tones, is a rather abstract concept. It is hard to define, and indeed it seems to be “impossible to find a (simple) acoustic correlate of stress” (Dutoit 1997:131). It causes prominence in syllable, but what exactly causes this prominence is a bit unclear. However, even if we cannot always produce it correctly, it is “a necessary concept for explaining the relation between prosodic structure (such as elementary contours) and segmental structure (such as word stress)” (Dutoit 1997:131).

Árnason (1998) shows that while lexical items in Icelandic are prosodically left-headed in Icelandic, neutral sentences are right-headed. That is to say, word level prosodic events generally occur on the first syllable, while sentence level prosodic events on the last word. He also states that “the stress goes on the rightmost ‘stressable’ unit of the focus domain” (Árnason 1998:50) Example (2.9), also from Árnason (1998), shows how neutral sentence stress is right strong.

Þetta er gamall MAður

'This is an old man' (2.9)

So it is the first syllable of the last word that receives the stress. This is not always the case as can be seen in example (2.10), once again from Árnason (1998:50).

Jón BAUð mér

'John invited me'

Jón bauð SIGgu

'John invited Sigga' (2.10)

There seems to be some sort of hierarchy in effect, and the above examples show how verbs are stronger than personal pronouns, while nouns are stronger than verbs. Árnason shows more examples of this, suggesting the following hierarchy:

nouns > verbs > prepositions > personal pronouns (2.11)

This seems to hold for many cases, but there are numerous exceptions, mostly having to do with semantics, and are therefore hard to predict automatically. When it comes to adverbials, the word class alone is not enough to determine the strength of the words as can be seen in (2.12) (Árnason, 1998).

Jón KEmur ekki

‘John isn’t coming’

Jón kemur Áreiðanlega

'John’s certainly coming’ (2.12)

Árnason goes into much more detail, but for automatic prosody generation we need simple rules we can follow, and armed with the knowledge we have already acquired, we should be able to make a system that predicts correct stress at least some of the time. As I have already mentioned, too many of the intricacies of prosody are beyond the rather poor analytical qualities of modern day synthesisers, and so we must make do with the limited information that is useable.

Duration plays some part in stress, as sounds in stressed syllables generally seem to be longer than in unstressed syllables. Duration therefore seems to affect “units” rather than single phonemes and Dutoit (1997:163) suggests a few formulas to calculate the duration of the phonemes within each unit. One of those is shown in (1.13).

$$Dur_i = exp(\mu_i + k\sigma_i) \quad (2.13)$$

Here, “Dur_i is the duration of the phoneme *i* in a syllable and μ_i and σ_i are the mean and standard deviation of its log-transformed durations in a large corpus” (Dutoit 1997:163). Using this formula, every phoneme within the same segment or unit will have its length factored by *k*, and *k* can be acquired from the corpus. Pending more research however it is hard to say how effective this formula will be.

2.5.2.2 Tones

Árnason (1998) mentions two contour tones in Icelandic, HL and LH (high-low and low-high), and two boundary tones, H% and L% (high and low). Recent research by Dehé (unpublished) seems to confirm this. These are usually used to contrast declarative and interrogative sentences, as seen in (2.14) (Examples from Árnason (1998)).

Þarna er DÍsa komin.

HL L%

‘There’s Dísa arrived’

Er DÍsa komin?

LH L%

‘Has Dísa arrived? (2.14)

Both interrogative and declarative sentences can end with a low boundary tone (L%), as well as a high tone (H%). These tones seem to signal finality or lack thereof as demonstrated in the correspondence between father, E, and his one-and-a-half-year-old son, V, in (2.15).

E: *Fyrst förum við í peysuna…*
HL H%
'First we put on the sweater...'
V: *Jáá…*
HL H%
'Yeeees...'
E: *Svo förum við í skóna…*
HL H%
'Then we put on the shoes…'
V: *Jáá…*
HL H%
'Yeees…'
E: *Og svo förum við út!*
HL L%
And then we go outside!
V: *Já!*
HL L%
Yes! (2.15)

In the first two utterances we have a rising boundary tone, and the child imitates this. The last utterance then ends with a falling tone, signalling the end of the series of actions, and again the child imitates this. This is not only an example of how we indicate whether there is more information to follow or not, but also shows that prosody is an important aspect of speech that is picked up by the child very early in the language acquisition process.

2.5.2.3 Phrases

Before looking at how tonal contours can be imported into a speech synthesiser, there is one last thing to look at. Prosody also has the function of segmenting the speech into shorter, more easily parsable parts. These are called minor prosodic phrases. The tones usually occur within those phrases, ending with a boundary tone. Dutoit (1997)

reports that the **chinks 'n' chunks algorithm** has proven very successful for right headed languages.

a (minor) prosodic phrase = chinks chunks** (2.16)

Chinks are basically defined as function words (prepositions and conjunctions) while chunks are content words. There are some important exceptions: only object pronouns are seen as chunks and tensed verb forms are considered chinks. For Icelandic, considering all verbs and all pronouns as chinks worked fairly well. There is little doubt that making the same distinction as Dutoit mentions, or finding new ones more suited for Icelandic, would improve the segmentation.

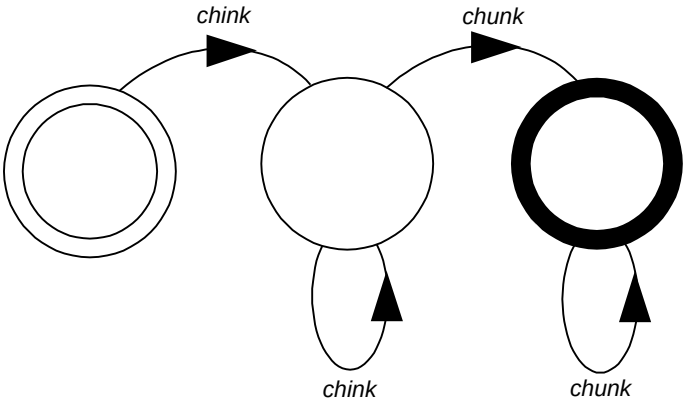

Figure 3 An FSA for the simple yet effective chinks 'n' chunks algorithm

While the segmentation into minor phrases using chinks ‘n’ chunks has shown some promise, it is still a very simple method that can easily go wrong. Consider the following sentence, taken from the RÚV corpus.

**(Eins) (og flestir á torginu) (bera Martin) (og Stefan húfur [..]) (í dönsku fánalitunum)*

‘Like most people on the square, Martin and Stefan wear hats in the Danish colours’ (2.17)

The error in (2.17) is the splitting of *eins og*, ‘like’, between two phrases. Whether this is something that should be accounted for by the POS-tagger or chinks ‘n’ chunks, an exception dictionary seems to be the only way to avoid this type of problem. It should be noted that a simple list of exception will hardly be enough as cases like (2.18) show.

(Jóhanna) (er tuttugu) (og eins) (og Guðrún) (er tuttugu) (og fimm.)

‘Johanna is twenty one and Gudrun is twenty five.’ (2.18)

In this case *eins og* does not mean ‘like’, but are indeed separate units, *eins* being the genitive neuter of ‘one’ and og being the conjunction ‘and’. Of course this

segmentation is not entirely correct, as numbers should probably be considered to be part of the same phrase: "*(Jóhanna)(er tuttugu og eins)*…" and so on.

Another issue not accounted for by chinks 'n' chunks are the differences between phrase boundaries. Simple rules for determining those differences are hard or impossible to find. Quené and Kager (1989) suggest distinguishing between **intonational phrases** and **phonological phrases**, and using their boundaries as guidelines.

[(This sentence) (is divided) (into intonational phrases)] [(which are divided) (into phonological phrases)]

(from Dutoit 1997:151) (2.19)

There is little data available for Icelandic phrase structure and this would be an interesting subject for further research, and would again be greatly aided by a spoken language corpus.

The nature of different boundaries should also shed some light on **downstep** in Icelandic, where the f0 frequency gradually decreases through each sentence.

Figure 2.4 Pitch contours of an utterance in newsreading. "*Halldór Ásgrímsson utanríkisráðherra er sömuleiðis staddur í Istanbúl.*" The two straight lines emphasise the gradual downstep.

Downstep does not occur in every utterance, and Árnason (1998) puts this down to a difference in finality and non-finality. When counting, for instance, downstep is less likely to occur.

Einn, tveir, þrír, fjórir…

LH LH LH LH

'One, two, three, four…' (2.20)

Each of the LH tones has roughly the same final pitch target. But in what in what Árnason (1998) calls "closed" counts, downstep can indeed occur.

Einn, tveir, þrír, fjórir, fimm.

LH LH LH LH HLL%

‘One, two, three, four, five. (2.21)

Here each LH tone has a slightly lower pitch target than the previous one. Both types of counting end with a final low tone eventually, however, but the downstep only seems to occur where the final number is certain from the beginning.

2.5.2.4 Speech Synthesis

I will be suggesting some rules that might work for predicting intonation, but such heuristics based rules made by hand will never compete with more advanced corpus based techniques.

It seems that for neutral speech synthesis the combinations HL, HLH%, HLL% and LHL% will suffice for the most part. Every phonological phrase will have the pattern HL, except the last phonological phrase within an intonational phrase, which will have the pattern HLH%, until the final phrase where HLL% will be used for declarative sentences, or LHL% for interrogative sentences. While the LH tone pattern does have a semantic and stylistic significance, this will not be covered by this system to ensure neutrality of the intonation and for the sake of simplicity. As Dehé (unpublished) states, both LH and HL are used to mark narrow focus (both contrastive and non-contrastive). HL however is more common of the two and so that will be our pitch accent of choice. The pseudo-code in (2.22) has given some promising results. Note that all pitch changes are gradual.

1. Define variables. PhraseInitialPitch = 140 Hz. StressPitch = 180 Hz. MinimumPitchForReset = 110 Hz. EndPitch = 75 Hz
2. Set pitch at the first syllable to PhraseInitialPitch.
3. If the phrase is the last phrase in the sentence:
    a. If the sentence is declarative: Set pitch at the penultimate syllable as StressPitch. Set the final syllable pitch to start at PhraseInitialPitch – 10 Hz and end at EndPitch. End program.
    b. If the sentence is interrogative: Set pitch at the penultimate syllable to start at PhraseInitialPitch – 10 Hz and end at StressPitch. Set the final syllable pitch to start at PhraseInitialPitch and end at EndPitch. End program.
4. Else: Set pitch at the penultimate syllable to start at StressPitch. Set the final syllable pitch to start at PhraseInitialPitch – 10Hz and end at StressPitch. Subtract 10 Hz from StressPitch, 5 Hz from PhraseInitialPitch.
    a. If PhraseInitialPitch is less than MinimumPitchForReset: go to 1.

b. If next phrase is in another intonational group: go to 2.

c. If next phrase is in same intonational group: go to 3. (2.22)

As an example, I selected a sentence from the RÚV corpus: *Franski ökuþórinn Sebastian Loeb, sem ekur Citroen-bíl, sigraði í Tyrklandsrallinu í morgun og náði níu stiga forystu í heimsmeistarkeppninni.* (‘French driver Sebastian Loeb, who drives a Citroen, won the Turkish rally this morning and now has a nine point lead in the world championship’).

[(Franski ökuþórinn Sebastian Loeb,)][(sem ekur Citroen-bíl,)][(sigraði í Tyrklandsrallinu í morgun) (og náði níu stiga forystu) (í heimsmeistarakeppninni.)] (2.23)

Chinks ‘n’ chunks should be able to determine the minor phrases correctly in this case. Accurate rules for determining intonational phrases are missing as I mentioned above, but in this case punctuation would have helped. Then we apply the rules for modifying f0 mentioned above. At the start the pitch is set at 140 Hz and gradually gets higher before reaching 180 Hz at the penultimate syllable. Then it drops down to 130 Hz (PhraseInitialPitch - 10 Hz), and at the start of the last syllable (Loeb), starts moving up again towards 180 Hz. We now detract 10 Hz from StressPitch and 5 Hz from the PhraseInitialPitch and start again with the second phrase. At the start the pitch is set at 135 Hz, moves up to 170 Hz at the penultimate syllable before dropping to 125 Hz and then going up to 170 Hz again. And so on.

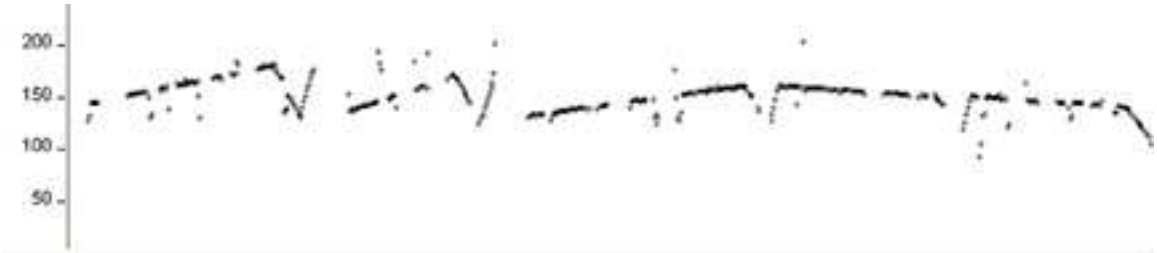

Figure 2.5 Pitch plot for a neutral synthesised sentence.

This sounds fairly natural already, even though the stress seems to be a bit off. Note how detracting from the PhraseInitialPitch and StressPitch after every phrase results in a slight downstep. In longer sentences, the downstep will be reset after PhraseInitialPitch reaches a certain minimum, defined by MinimumPitchForReset.

Adding stress information to the pseudo-code is not too hard. Simply moving the StressPitch position from the penultimate syllable to the syllable to be stressed seems to work fairly well. In the above example, moving the pitch peak in the second phrase

to the first syllable of ‘Citroen’ and the first syllable of ‘Tyrklandsrallinu’ in the third phrase and so on does seem to make the utterance sound significantly more natural.

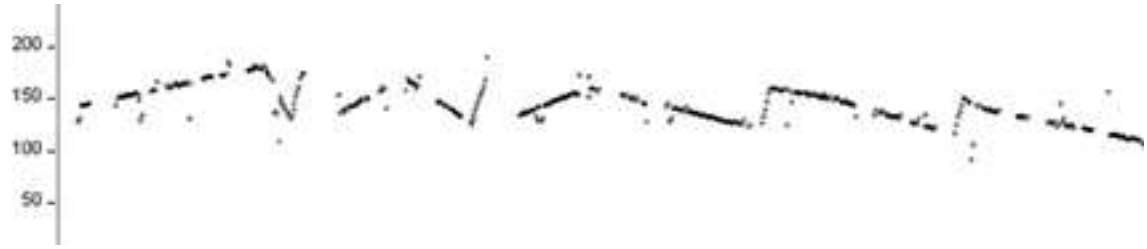

Figure 2.6 Stress adjusted synthesised speech.

But will this always work? The simple answer is no. And unless an acceptably accurate automatic way to do the phrasal segmentation can be found, the majority of sentences will sound wrong. It would be easier to allow the user to segment the text himself by hand, and indeed this is the solution many TTS systems use (Dutoit 1997:152). I do believe that the pseudo-code will work fairly well in many cases, provided the segmentation is correct. But manual segmentation is hardly acceptable, as one of the main strengths of having a synthesiser is its ability to read dynamic data. Automatically determining the segmentation correctly is beyond the current system, and with the exact nature of intonational and phononological phrases undetermined finding ways of overcoming this problem will be hard or impossible. Even if the segmentation was correct, in its current form the pseudo-code has some great flaws. Short phrases, for instance, would cause problems, especially exceedingly short phrases as those in (2.24).

[(Jón) (fór).])

‘John left.’ (2.24)

The pseudo code assumes at least two syllables per segment (StessPitch on the penultimate syllable, PhraseInitialPitch and EndPitch on the last syllable), so these have to be accounted for as well. And it can only read a specific type of text; to capture more intricacies of speech a large set of rules would be required. As before, automatic corpus based training must be the recommended option.

2.6 Icelandic Synthesis

It is my belief that language technology is something Iceland could become quite good at on an international scale. The equipment required does not have to be expensive, but the results can be highly profitable. Speech synthesisers are developing rapidly and speech synthesis solutions are being sought worldwide. Speech synthesis

in Iceland does not have to be only synthesis of Icelandic, and focusing on language independent solutions as much as possible will result in knowledge and experience that can be exported.

But we must not stretch ourselves too far too fast. Above I have mentioned some of the parts of speech synthesis where work is required. By using freely available systems such as the Edinburgh-made Festival system, which is already becoming standard-setting in the world of speech synthesis (Dutoit 1997), it is possible to experiment with the different modules separately and try them out as work progresses, rather than blindly hoping that it will all come together in the end. Festival even supports MBROLA databases, so the first steps have already been taken. Even though the MBROLA database must not be used for commercial purposes, it can be used as a placeholder for later recording while experimenting with every aspect of the narrow phonetic transcription.

For the different parts of speech synthesis I have covered so far, I hope I have managed to show both options available with what knowledge and data is available now, but also how what steps can be made in the near future.

Those future steps all depend on a large and well-made corpus of Icelandic speech, and how this can be acquired will be the subject of the next chapter.

3. Building a Phonetic Corpus

Most of what I have said above is based on supposition rather than empirical evidence. Indeed, on top of that much of the recommended methodology for automatic generation of prosody and phonetisation presupposes the existence of a large phonetic and prosodic corpus. I therefore must suggest that the first step taken towards high quality emotive speech synthesis for Icelandic must be the building of such a corpus. Work on a basic corpus of this nature is already well underway, and in the following chapters I shall be describing the work completed so far, what must be improved and suggest further development required for a full fledged phonetic and prosodic corpus of Icelandic.

3.1 (vef)Setur hljóðan(na)

To assist prosodic research for Icelandic speech synthesis an experimental corpus was created to gain some further practical understanding of what to consider when building a phonetic and prosodic corpus. It was given the name (vef)Setur hljóðan(na) (a play on words that is hard to translate), and it can be found at http://hljodan.simblogg.is. The corpus is composed entirely of recordings of news reading provided by the Icelandic National Radio (RÚV). No interviews are included, only read material. This provides a good example of 'professional reading', something that might be a desirable target for a speech synthesis system. It has its limitations, however, as the reading is not very emotional, and there is little variation in the types of sentences. For instance there was a single question in the whole of the corpus.

The recordings were copied from CDs and converted into wav files, each containing a sentence or utterance. For determining where an utterance ended breaks in speech after a fall in tone were used rather than for instance punctuation.

A part of the corpus was transcribed phonetically and further analysed using MBROLIGN for automatic phonetic alignment and praat for fine-tuning the alignment and extracting data. Mietta Lennes' *make_textgrid_from_segment_data*[11] program for praat was invaluable for translating the MBROLIGN data to praat readable textgrids. Praat is a powerful program for doing phonetic analysis and can be downloaded for various operating systems free of charge from the following website: http://www.fon.hum.uva.nl/praat/

[11] http://www.helsinki.fi/~lennes/praat-scripts/.

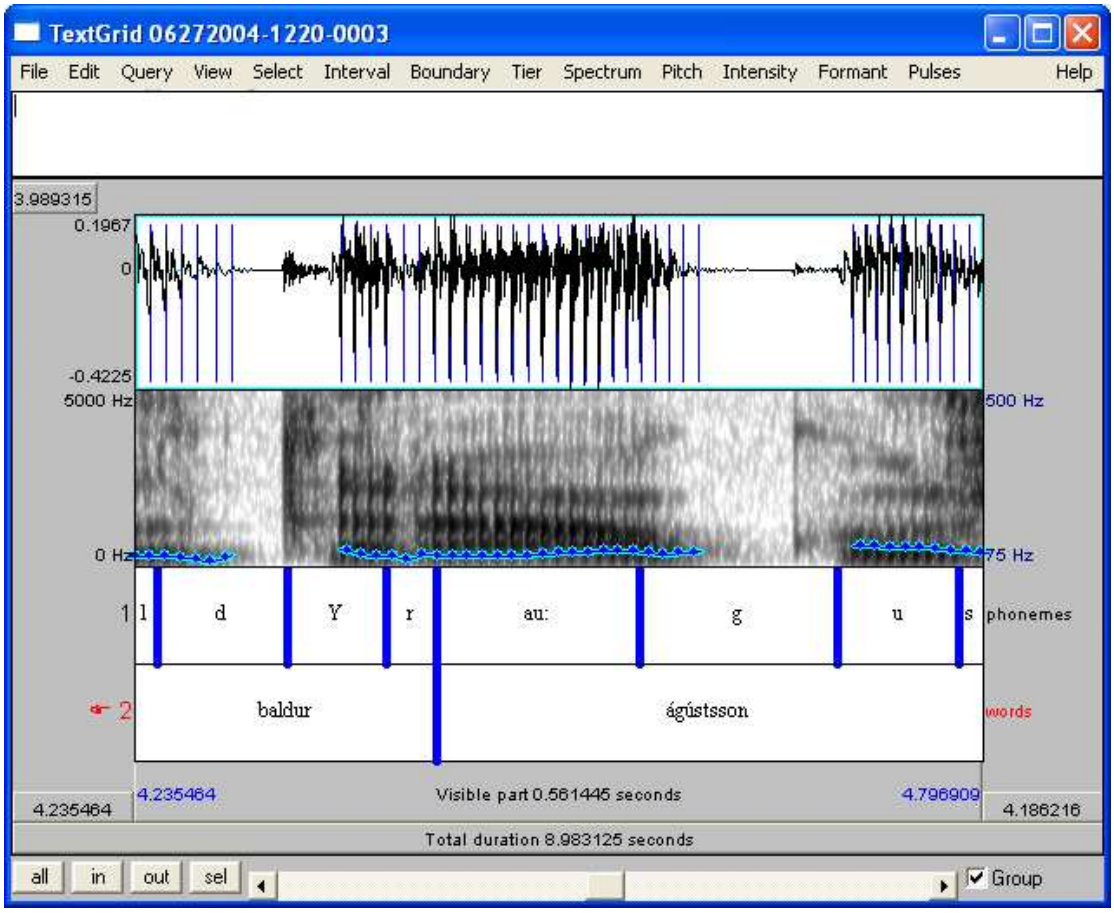

Figure 3.1 A praat textgrid showing waveform, spectrogram, phonetic and orthographic transcription tiers.

When part of the corpus had been aligned like this, with only phonetic transcription and waveforms to match, all sorts of information could be extracted. Praat has its own scripting language so it could be programmed to extract every occurrence of a phone or diphone, which could help with locating potential targets for diphone system creation. It can also be used to find highest pitch of a phrase to determine not only where rises occur and when during the phone or syllable they peak, but also give numerical data to show exactly how high the pitch is. These sorts of data are doubtless of great value as guidelines for prosodic generation until fully automatic methods can be developed. With sophisticated praat scripts like Piet Mertens' prosogram[12] a detailed graphical analysis, again, using only a sound file and a textgrid as data, can be generated (fig. 3.2). A graphical representation of sound files like this could be very helpful to transcribers looking for boundaries and locations of tones.

[12] http://bach.arts.kuleuven.ac.be/pmertens/prosogram/

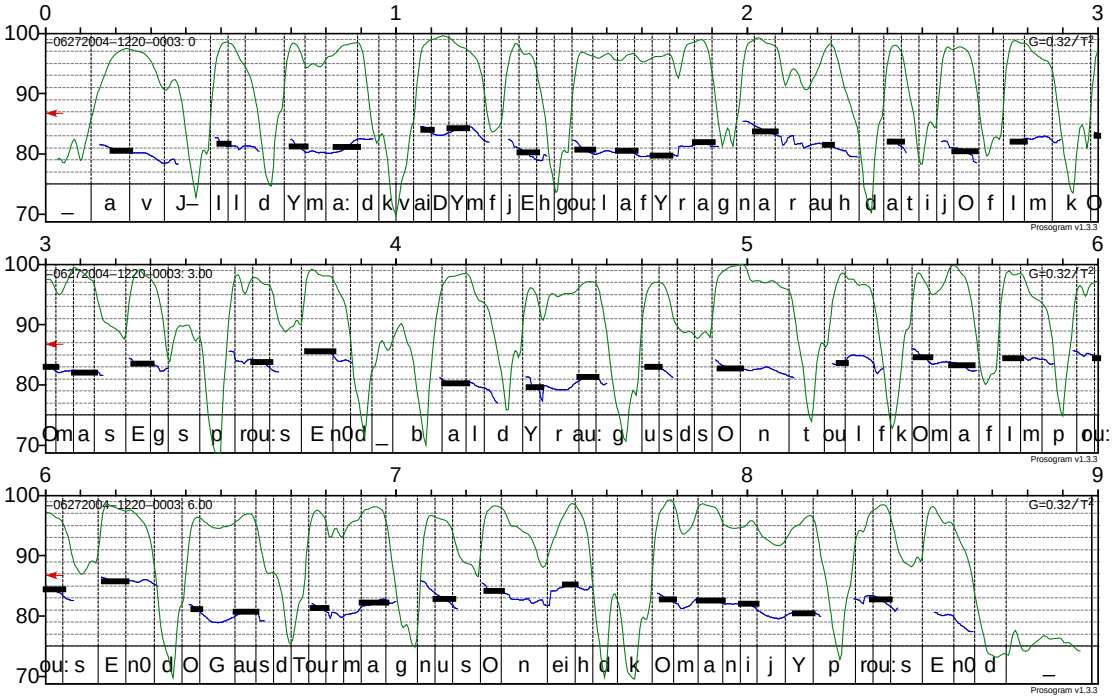

Figure 3.2 A prosogram graph showing pitch, power and phonetic transcription. Please note that this is using the older SAMPA standard.

The text was automatically tagged using a TnT based POS tagger being developed by Sigrún Helgadóttir et al. The tagging was minimal, only the word class was determined, but details such as case, number and gender were ignored.

The corpus was placed on the web for public use, and for navigation a front end was programmed in php. The corpus structure is simple: for each news hour the sound files have a unique prefix, followed by an incremental number. For each news hour there is also a text file with the same prefix, and the contents of the text file have a one-to-one correspondence with the sound files. That is to say, line one of the text file describes sound file one, line two goes with sound file two and so on. The text files describe what is being said, but also the word class of each word as determined by the POS tagger. In the corpus, the words are colour coded based on the word class. A link to the corresponding sound file is also provided. Furthermore, the user has the option to display automatically determined intonation phrases, based on the *chinks 'n' chunks* algorithm described above, for comparison.

Great emphasis has been placed on keeping the content separate from the design, making it easy it import more data into the corpus without any additional programming. So far it has succeeded in my view. Of course any annotational additions would have to be programmed into the corpus, but adding more data of the

current form would only require the sound files and the corpus description in a text file to be uploaded to the site. This is important for future changes.

But in the end the (vef)Setur hljóðan(na) corpus must be considered a pilot corpus, an experiment to see what can be done and how it should be done, a small project to learn about the 'dos and don'ts' before undertaking a larger project. This is what I hope can be done.

3.2 Towards a Spoken Language Corpus

Making the corpus will be no trivial task, and must be planned extensively beforehand to ensure a successful outcome. Hopefully the following guidelines will provide a good starting point.

3.2.1 Data Storage

In recent years storing speech recordings digitally has become the norm and must be considered the preferred and obvious option. But what happens when data are converted from analogue to digital? Understanding this is important so the recordings can be of the highest quality possible, while keeping the size manageable. Johnson's (2003) excellent analysis is the basis for what follows on digital storing of sound.

3.2.2.1 The Nature of Sound

First we must understand what sound is and what we are recording.

> A sound wave is a travelling pressure fluctuation that propagates through any medium that is elastic enough to allow molecules to crowd together and move apart. The wave in a lake after you throw in a stone is an example. The impact of the stone is transmitted over a relatively large distance. The water particles don't travel; the pressure fluctuation does. (Johnson 2003:4)

These fluctuations can be measured in **cycles per second (Hz)** where the higher the frequency (again, Hz), the higher the pitch.

3.2.2.2 Analogue and Digital

To store data digitally we have to split the sound into samples and store those samples individually. Sound waves are usually represented by a curve: an analogue recording would best be described by a solid line (a continuous wave), while a digital recording is a collection of points, each point representing a sample (a discrete wave).

The frequency of the recording is a measurement of how many of these samples are played in a second. The more samples we get, i.e. the higher the frequency of the recording, the better the quality. But how far can we go? If we look at a simple sinus wave, we can see that we need at least two points to store it: one for the highest point

and another for the lowest valley (assuming the frequency remains constant). This is the minimum, having more points would of course increase the accuracy of the recording. If we keep this in mind, we can see that because the frequency of sound is measured in cycles per second and the frequency of a recording is measured in samples per second, the highest frequency sound we can capture is half that of the frequency of the recording. To reiterate: to record a cycle we need at least two samples, therefore to record a sine wave that repeats 100 times per second (100Hz), we need 2 samples to record a single cycle, and therefore a recording frequency of 200Hz. This is why CDs are recorded at 44kHz, as the human ear is incapable of hearing sounds of a frequency higher than about 20kHz. There is a term for this; the highest frequency that can be captured with a given sampling rate is called a Nyquist frequency (Johnson 2003:22).

Figure 3.3 To represent a constant sine wave, at least two samples are required for each cycle.

Another factor in the quality of a recording is its bit rate. The bit rate determines the accuracy of the samples. With more bits we can use more digits to represent each sample. In an 8-bit recording, for instance, we only have 256 values to choose from and this will give us a 'stepped' wave as can be seen in (Fig. 3.4). It will not conform very well to the original. This deviation from the original sound wave generates noise. The higher the bit rate, the more accurately the digital representation can approximate the original sound recording and the lower the noise level. CDs, for instance, use 16 bit recordings. There will be some stepping in every digitised sound wave because we are representing a continuous wave as a discrete one, but with a higher bit rate steps will be smaller and in greater numbers, making the deviation so minimal that it is hardly audible.

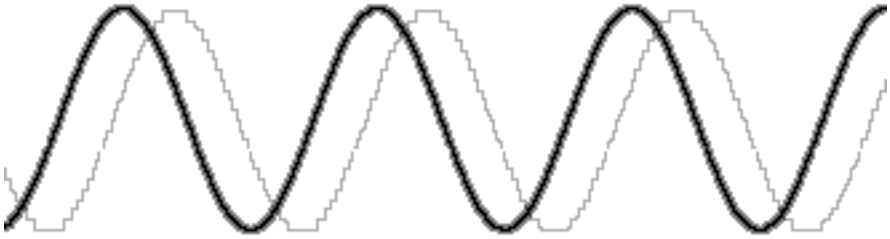

Figure 3.4 Two identical sinus waves recorded at 44 kHz, once with a bitrate of 16 (dark line) and again at a bitrate of 8 (light line). The waves have been slightly offset to make the difference clearer.

3.2.2.3 Excess Frequencies

This is not the end of the matter though, as the sounds of a frequency higher than can be recorded at a certain frequency are not ignored automatically. The computer will try and record them anyway, and tries to generate a wave from the points it can 'see'. This results in a wave of a frequency much lower than the original and causes noise or distortion in the sound file. Therefore, the higher frequency sounds must be filtered out.

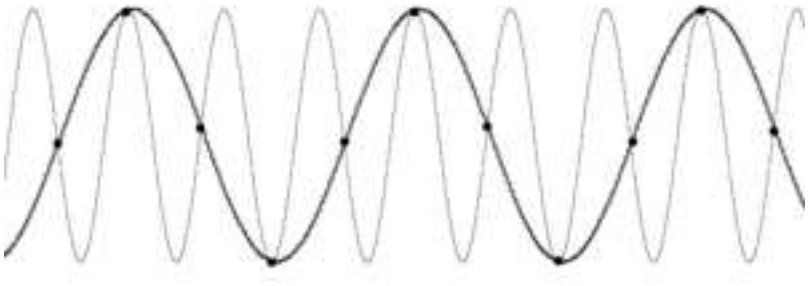

Figure 3.5 A high frequency wave (light) cannot be properly described with a lower frequency recording. The resulting wave (dark) is of a much lower frequency than the original.

This is perhaps why some insist that vinyl records are superior to digital recordings, even though the lost sound is of a frequency that should be completely out of range for the human ear.

3.2.2.4 Storing Data for Speech Research

In speech, frequencies above 10kHz are not likely to be of any consequence at all to speech research so this should not be a worry, and storing data at 22kHz should be more than sufficient. In the past, 8 bit recordings were considered good enough, or at least preferable because of size limitations, as the recordings would have to be stored on floppies or transmitted through much slower Internet connections, but nowadays file size is the least of our worries. However, there is no need to go overboard and 16 bit 22 kHz recordings should store all the data that are needed accurately.

3.3 Collecting the Data

Having decided on the format of the data, a decision on what should be recorded must be taken. The corpus must be usable for the purposes mentioned above, but it should also have some potential for the future. What could be used right now would be data for phonetisation and prosodic training. A database for diphones would also be handy to improve and replace the MBROLA database, but as mentioned above, diphone databases are highly specialised and recording them is not all that time consuming. When carefully prepared, all the recordings can be made in an afternoon and the segmentation in a couple of days. But looking to the future, a prosodically and phonetically annotated corpus is a large step in the direction of unit selection synthesis, so while such a corpus would be much bigger than what is strictly required for a diphone synthesiser, it would still be very useful, and invaluable for future synthesis.

3.3.1 What to Record and How Much

It is hard to say exactly how large a corpus would have to be to function as a database for a unit selection system, but Black and Lenzo (2003) provide some guidelines. A phonetically balanced database of 460 sentences is “not unreasonable”, while “[i]f the text has not been specifically selected for phonetic coverage a larger database is probably required”. As the aim of the corpus is quite clearly defined, it should be fairly well balanced phonetically, and therefore 460-500 sentences should be enough.

Black and Lenzo (2003) suggest that “If the database is broadcast news stories, the synthesis from it will typically sound like read news stories (or more importantly will sound best when it is reading news stories).” With this in mind it is important to select the domain carefully. So far, speech synthesis has been mostly mentioned in the context of phone services. While it is likely that such a synthesiser will be required to do some news reading, a slightly more lively personality might be better. This should be acquired by recording various common greetings and welcome messages and other such social utterances, along with a few possible service options like “would you like…” and “might I offer you a…” and so on. Along with news texts, something like text for tourists might be fitting.

3.3.2 Labelling

Compared to the all-important labelling of the corpus, the recording process is a trivial matter. The labelling will consist of aligning words, phonemes, prosodic tones and boundaries with the waveform, and for a large corpus this is extremely tedious and time consuming. Fortunately there are ways of doing this automatically using ASR technology. Tools for this purpose are freely available. Those are mostly multi-lingual and it might take some time to adapt them to Icelandic, so finding out if the new Icelandic ASR engine could be of any use might be worth the effort. These tools are usually fairly robust however, and tools like Carnegie Mellon's SphinxTrain along with the free Sphinx speech recognition system have "been reliably used to labeling [sic] hundreds of databases in many different languages" (Black and Lenzo 2003).

Automatic labelling will always be inadequate however, and manual corrections will always be required. This is especially true of phonetic alignment. This is for the reason that ASRs are based on phoneme centres and mostly ignore phoneme boundaries. Speech synthesis, however, is based on phoneme boundaries and so it is more important to get the boundaries accurately rather than finding the centre. The result may often be quite close, but if the boundaries are only a few milliseconds off, this can have disastrous results for any speech synthesisers made from data extracted from the corpus, as stray sounds will enter into the diphone and interrupt the flow of the speech.

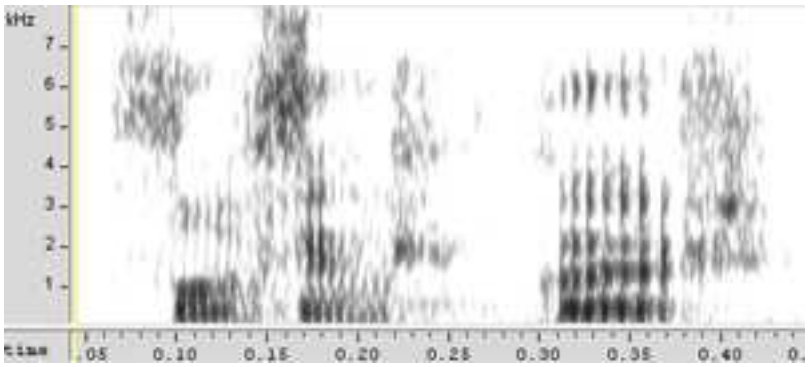

Figure 3.6 A misaligned diphone. The word is *þykkur*, ('thick'), but the second diphone [T I] has a part of an [u] at its start. Without knowing that the word should be *þykkur*, it is nearly impossible to understand.

3.4 Using the Corpus

A corpus that is never used is not worth much. While it can be hard to predict how useful a corpus will be beforehand, it goes without saying that if the corpus is hard to access it is hardly going to be much good to anyone. I hope I have shown that a

phonetic corpus for Icelandic would be of immense value for Icelandic speech research. Making it publicly available online is one way of making it available, even if only part of it is. It may be hard to adapt all the annotated data reliably to the web format, but even if only the sounds along with a aligned and searchable orthographic transcription were available, this might be enough for the casual user, and public enough so the more serious researcher would know these data existed and could get access to the more specialised labelling along with the required software without too much trouble.

4. Final Words

I have no doubt that the knowledge required to make speech synthesis systems in Iceland already exists. The data, however, are missing, but these can be acquired and put to use. With proper funding, a decent corpus and a freely available and fully working Icelandic speech synthesiser could be available in just over a year. With ongoing work the corpus can be developed further, and high quality data driven synthesis could become a viable option. With even further development I could see data driven synthesisers for other languages being exported from Iceland within the next few years.

Works Cited

Árnason, Kristján. 1998. Toward an analysis of Icelandic intonation. Stefan Werner (ed.) *Nordic Prosody VII*, pp. 49-62. Peter Lang, Frankfurt am Main, Germany.

Burileanu, Dragos. 2002. Basic Research an Implementation Decisions for a Text-to-Speech Synthesis System in Romanian. *International Journal of Speech Technology 5*, pp. 211-225. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Bigorne, D., O. Boeffard, B. Cherbonnel, F. Emerard, D. Larreur, J.L. le Saint-Milon, I. Metayer, C. Sorin and S. White. 1993. Multilingual PSOLA Text-to-Speech system. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing 93*, vol 2, pp. 187-190. Minneapolis, USA.

Black, Alan W. and Kevin A. Lenzo. 2003. *Building Synthetic Voices*. <http://www.festvox.org/bsv/>

Carlson, Rolf and Björn Granström. 1997. Speech Synthesis. W. Hardcastle and J. Laver (eds.): *The Handbook of Phonetic Sciences*, pp. 768-788. Blackwell Publishers Ltd, Oxford, England.

Dehé, Nicole. Unpublished. Some Notes on the Focus-Prosody Relation and Phrasing in Icelandic. To appear in Gösta Bruce and Merle Horne (eds.): *Nordic Prosody IX*.

Dutoit, Thierry. 1993. *High Quality Text-to-Speech Synthesis of the French Language*, Ph.D. dissertation, Faculté Polytechnique de Mons, Belgium.

Dutoit, Thierry. 1997. *An Introduction to Text-to-Speech Synthesis.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

Fordyce, Cameron S. 1998. *Prosody for Speech Synthesis using Transformational Rule-Based Learning.* Boston University, College of Engineering, Boston, USA.

Helgason, Pétur. 1990. *Lokaskýrsla verkefnis um tölvutal*. Málvísindastofnun HÍ, Reykjavík, Iceland.

Hirschberg, Julia. 2002. Communication and prosody: Functional aspects of prosody. *Speech Communication: Special Issue on Dialogue and Prosody,* pp. 31-43. J. Terken and M. Swerts (eds.).

Holmes, John and Wendy Holmes. 2001. *Speech Synthesis and Recognition*. 2nd ed. Taylor and Francis, London, England.

Johnson, Keith. 2003. *Acoustic & Auditory Phonetics*. 2nd ed. Blackwell Publishing Ltd, Oxford, England.

Lenzo, Kevin A. 1998. *s/($text)/speech $1/eg;*. <http://www-2.cs.cmu.edu/~lenzo/tpj/tpj12_synthesis.html>.

Monaghan, A.I.C. 1990. A multi-phrase parsing strategy for unrestricted text. *Proc. ESCA Workshop on speech synthesis*, pp. 109-112. Autrans, France.

Pétursson, Magnús. 1974. Recherche expérimentale sur le problème de la quantité en islandais moderne et sur les caracteristiques secondaires des voyelles islandaises. *Travaux de l'Institut de Phonétique de Strasbourg* 6:23-64.

Rögnvaldsson, Eiríkur. 1989. *Íslensk hljóðfræði*. Málvísindastofnun Háskóla Íslands, Reykjavík, Iceland.

Rögnvaldsson, Eiríkur. 2003. *Phonetic Transcription Guideline: Icelandic*.

Syrdal, Ann K., C.W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K-S. Lee, M.J Makashay. 2000. Corpus-Based Techniques in the AT&T Nextgen Synthesis System. *Proc. of ICSLP 2000, vol. III*, pp. 410-415. Beijing, China.

Appendix

Icelandic Diphones

These are the diphones that were recorded for the MBROLA synthesiser, adapted to the new SAMPA standard.

6y	C	6y:	l0	9	i
6y	D	6y:	m	9	i:
6y	G	6y:	n	9	j
6y	J	6y:	n0	9	k
6y	J0	6y:	p	9	k0
6y	N	6y:	p0	9	l
6y	N0	6y:	r	9	l0
6y	T	6y:	r0	9	m
6y	_	6y:	s	9	m0
6y	c	6y:	t	9	n
6y	c0	6y:	t0	9	n0
6y	f	6y:	v	9	ou
6y	h	9	6y	9	ou:
6y	j	9	6y:	9	oy
6y	k	9	9	9	p
6y	k0	9	9:	9	p0
6y	l	9	A	9	r
6y	l0	9	A:	9	r0
6y	m	9	C	9	s
6y	m0	9	D	9	t
6y	n	9	E	9	t0
6y	n0	9	E:	9	v
6y	p	9	G	9	x
6y	p0	9	I	9:	6y
6y	r	9	I:	9:	9
6y	r0	9	O	9:	A
6y	s	9	O:	9:	C
6y	t	9	T	9:	D
6y	t0	9	U	9:	E
6y	v	9	U:	9:	G
6y	x	9	Y	9:	I
6y:	D	9	Y:	9:	O
6y:	G	9	_	9:	T
6y:	T	9	au	9:	U
6y:	_	9	au:	9:	Y
6y:	c	9	ay	9:	_
6y:	c0	9	ay:	9:	au
6y:	f	9	c	9:	ay
6y:	h	9	c0	9:	c
6y:	j	9	ey	9:	c0
6y:	k	9	ey:	9:	ey
6y:	k0	9	f	9:	f
6y:	l	9	h	9:	h

9:	i	A	f	A:	l0
9:	j	A	h	A:	m
9:	k	A	i	A:	n
9:	k0	A	i:	A:	n0
9:	l	A	j	A:	ou
9:	l0	A	k	A:	oy
9:	m	A	k0	A:	p
9:	n	A	l	A:	p0
9:	n0	A	l0	A:	r
9:	ou	A	m	A:	r0
9:	oy	A	m0	A:	s
9:	p	A	n	A:	t
9:	p0	A	n0	A:	t0
9:	r	A	ou	A:	v
9:	r0	A	ou:	A:	x
9:	s	A	oy	C	9
9:	t	A	p	C	9:
9:	t0	A	p0	C	A
9:	v	A	r	C	A:
9:	x	A	r0	C	E
A	6y	A	s	C	E:
A	6y:	A	t	C	U
A	9	A	t0	C	U:
A	9:	A	v	C	Y
A	A	A	x	C	au
A	A:	A:	6y	C	au:
A	C	A:	9	C	ey
A	D	A:	A	C	ou
A	E	A:	C	C	ou:
A	E:	A:	D	D	6y
A	G	A:	E	D	6y:
A	I	A:	G	D	9
A	I:	A:	I	D	9:
A	J0	A:	O	D	A
A	O	A:	T	D	A:
A	O:	A:	U	D	C
A	T	A:	Y	D	E
A	U	A:	_	D	E:
A	U:	A:	au	D	I
A	Y	A:	ay	D	I:
A	Y:	A:	c	D	O
A	_	A:	c0	D	O:
A	au	A:	ey	D	T
A	au:	A:	f	D	U
A	ay	A:	h	D	U:
A	ay:	A:	i	D	Y
A	c	A:	j	D	Y:
A	c0	A:	k	D	au
A	ey	A:	k0	D	au:
A	ey:	A:	l	D	au:

D ay
D ay:
D c
D c0
D ey
D ey:
D ey:
D f
D i
D i:
D j
D k
D k0
D l
D l0
D m
D n
D n0
D n0
D ou
D ou:
D oy
D p
D p0
D r
D r0
D s
D t
D t0
D v
D yy
E 6y
E 6y:
E 9
E 9:
E A
E A:
E C
E D
E E
E E:
E G
E I
E I:
E O
E O:
E T
E U
E U:
E Y
E Y:
E _
E au
E au:
E ay
E ay:
E c
E c0
E ey
E ey:
E f
E h
E i
E i:
E j
E k
E k0
E l
E l0
E m
E m0
E n
E n0
E ou
E ou:
E oy
E p
E p0
E r
E r0
E s
E t
E t0
E v
E x
E: 6y
E: 9
E: A
E: C
E: D
E: E
E: G
E: I
E: O
E: T
E: U
E: Y
E: _
E: au
E: ay
E: c
E: c0
E: f
E: h
E: i
E: j
E: k
E: k0
E: l
E: l0
E: m
E: n
E: n0
E: ou
E: oy
E: p
E: p0
E: r
E: r0
E: s
E: t
E: t0
E: v
E: x
G 6y
G 6y:
G 9
G A
G A:
G C
G D
G E
G E:
G I
G I:
G O
G O:
G U
G U:
G Y
G Y:
G au
G au:
G ay
G ay:
G c
G c0
G ey
G ey:
G f

G	i	I	f	I:	n0
G	i:	I	h	I:	oy
G	j	I	i	I:	p
G	k	I	i:	I:	p0
G	k0	I	j	I:	r
G	l	I	j	I:	r0
G	m	I	k	I:	s
G	n	I	k0	I:	t
G	n0	I	l	I:	t0
G	ou	I	l0	I:	v
G	ou:	I	m	I:	x
G	oy	I	m0	J	c
G	p	I	n	J0	c
G	p0	I	n0	N	k
G	r	I	ou	N	l
G	t	I	ou:	N	n
G	t0	I	oy	N	p
G	v	I	p	N0	k
I	6y	I	p0	N0	t
I	6y:	I	r	O	6y
I	9	I	r0	O	6y:
I	9:	I	s	O	9
I	A	I	t	O	9:
I	A:	I	t0	O	A
I	C	I	v	O	A:
I	D	I	x	O	C
I	E	I:	6y	O	D
I	E:	I:	9	O	E
I	G	I:	A	O	E:
I	I	I:	C	O	G
I	I:	I:	D	O	I
I	J	I:	E	O	I:
I	N	I:	G	O	N
I	N0	I:	I	O	O
I	O	I:	O	O	O:
I	O:	I:	T	O	T
I	T	I:	U	O	U
I	U	I:	Y	O	U
I	U:	I:	_	O	U:
I	Y	I:	au	O	Y
I	Y:	I:	c	O	Y:
I	_	I:	c0	O	_
I	au	I:	f	O	au
I	au:	I:	h	O	au:
I	ay	I:	k	O	ay
I	ay:	I:	k0	O	ay:
I	c	I:	l	O	c
I	c0	I:	l0	O	c0
I	ey	I:	m	O	ey
I	ey:	I:	n	O	ey:

O f
O h
O i
O i:
O j
O k
O k0
O l
O l0
O m
O m0
O n
O n0
O ou
O ou:
O oy
O p
O r
O r0
O s
O t
O t0
O v
O x
O: 6y
O: 9
O: A
O: C
O: D
O: E
O: G
O: I
O: O
O: T
O: Y
O: _
O: au
O: ay
O: c
O: c0
O: ey
O: f
O: h
O: i
O: j
O: k
O: k0
O: l
O: l0
O: m
O: n
O: n0
O: ou
O: oy
O: p
O: r
O: r0
O: s
O: t
O: t0
O: v
T 6y
T 6y:
T 9
T 9:
T A
T A:
T C
T E
T E:
T I
T I:
T O
T O:
T U
T U:
T Y
T Y:
T _
T au
T au:
T ay
T ay:
T c
T c0
T ey
T ey:
T f
T h
T i
T i:
T j
T k
T n0
T ou
T ou:
T p
T p0
T r
T r0
T s
T t
T t0
T v
U 6y
U 6y:
U 9
U 9:
U A
U A:
U C
U D
U E
U E:
U G
U I
U I
U I:
U J
U J0
U N
U N0
U O
U O:
U T
U U
U U:
U Y
U Y
U _
U au
U au:
U ay
U ay:
U c
U c0
U ey
U ey:
U f
U h
U i:
U j
U k
U k0
U l
U l0
U m
U m0
U n

U n0
U ou
U ou:
U oy
U p
U p0
U r
U r0
U s
U t
U t0
U v
U x
U: 6y
U: 9
U: A
U: C
U: D
U: E
U: G
U: I
U: O
U: T
U: U
U: U
U: Y
U: _
U: au
U: ay
U: c
U: c0
U: ey
U: f
U: h
U: i
U: j
U: k
U: k0
U: l
U: l0
U: m
U: n
U: n0
U: ou
U: oy
U: p
U: p0
U: r
U: r0
U: s
U: t
U: t0
U: v
U: x
Y 6y
Y 6y:
Y 9
Y 9:
Y A
Y A:
Y C
Y D
Y E
Y E:
Y G
Y I
Y I:
Y O
Y O:
Y T
Y U
Y U:
Y Y
Y Y:
Y _
Y au
Y au:
Y ay
Y ay:
Y c
Y c0
Y ey
Y ey:
Y f
Y h
Y i
Y i:
Y j
Y k
Y k0
Y l
Y l0
Y m
Y m0
Y n
Y n0
Y ou
Y ou:
Y oy
Y p
Y p0
Y r
Y r0
Y s
Y t
Y t0
Y v
Y x
Y: 6y
Y: 9
Y: A
Y: C
Y: D
Y: E
Y: G
Y: I
Y: O
Y: T
Y: Y
Y: _
Y: au
Y: ay
Y: c
Y: c0
Y: ey
Y: f
Y: h
Y: i
Y: j
Y: k
Y: k0
Y: l
Y: l0
Y: m
Y: n
Y: n0
Y: ou
Y: oy
Y: p
Y: p0
Y: r
Y: r0
Y: s
Y: t
Y: t0
Y: v
Y: x
_ 6y
_ 6y:
_ 9

_	9:	au	A	au	x
_	A	au	A:	au:	6y
_	A:	au	C	au:	9
_	C	au	D	au:	A
_	E	au	E	au:	C
_	E:	au	E:	au:	D
_	I	au	I	au:	E
_	I:	au	I:	au:	G
_	O	au	J	au:	I
_	O:	au	J0	au:	O
_	T	au	N	au:	T
_	U	au	N0	au:	U
_	U:	au	O	au:	Y
_	Y	au	O:	au:	_
_	Y:	au	T	au:	au
_	au	au	U	au:	ay
_	au:	au	U:	au:	c
_	ay	au	Y	au:	c0
_	ay:	au	Y:	au:	ey
_	c	au	_	au:	f
_	c0	au	au:	au:	h
_	ey	au	ay	au:	i
_	ey:	au	ay:	au:	j
_	f	au	c	au:	k
_	h	au	c0	au:	k0
_	i	au	ey	au:	l
_	i:	au	ey:	au:	l0
_	j	au	f	au:	m
_	k	au	h	au:	n
_	k0	au	i	au:	n0
_	l	au	i:	au:	ou
_	l0	au	j	au:	oy
_	m	au	k	au:	p
_	n	au	k0	au:	p0
_	n0	au	l	au:	r
_	ou	au	l0	au:	r0
_	ou:	au	m	au:	s
_	oy	au	n	au:	t
_	p	au	n0	au:	t0
_	p0	au	ou	au:	v
_	r	au	ou:	au:	x
_	r0	au	oy	ay	C
_	s	au	p	ay	D
_	t	au	p0	ay	G
_	t0	au	r	ay	J
_	v	au	r0	ay	N
au	6y	au	s	ay	N0
au	6y:	au	t	ay	T
au	9	au	t0	ay	_
au	9:	au	v	ay	c

ay c0
ay f
ay h
ay j
ay k
ay k0
ay l
ay l0
ay m
ay m0
ay n
ay n0
ay p
ay p0
ay r
ay r0
ay s
ay t
ay t0
ay v
ay x
ay: C
ay: D
ay: G
ay: T
ay: _
ay: c
ay: c0
ay: f
ay: h
ay: j
ay: k
ay: k0
ay: l
ay: l0
ay: m
ay: n
ay: n0
ay: p
ay: p0
ay: r
ay: r0
ay: s
ay: t
ay: t0
ay: v
ay: x
c 9
c 9:
c A
c A:
c E
c E:
c I
c I:
c O
c U
c U:
c Y
c au
c au:
c ay
c ay:
c ey
c ey:
c i
c i:
c ou
c ou:
c0 9
c0 9:
c0 A
c0 A:
c0 E
c0 E:
c0 I
c0 I:
c0 O
c0 Y
c0 au
c0 au:
c0 ay
c0 ay:
c0 ey
c0 ey:
c0 i
c0 i:
c0 ou
c0 ou:
ey C
ey D
ey G
ey J
ey J0
ey N
ey N0
ey T
ey _
ey c
ey c0
ey f
ey h
ey j
ey k
ey k0
ey l
ey l0
ey m
ey m0
ey n
ey n0
ey p
ey p0
ey r
ey r0
ey s
ey t
ey t0
ey v
ey x
ey: C
ey: D
ey: G
ey: T
ey: _
ey: c
ey: c0
ey: f
ey: h
ey: j
ey: k
ey: k0
ey: l
ey: l0
ey: m
ey: n
ey: n0
ey: p
ey: p0
ey: r
ey: r0
ey: s
ey: t
ey: t0
ey: v
ey: x
f 6y
f 6y:
f 9
f 9:

f	A	h	E	i	r0
f	A:	h	E:	i	s
f	C	h	I	i	t
f	E	h	I:	i	t0
f	E:	h	O	i	v
f	I	h	O:	i	x
f	I:	h	U	i:	C
f	O	h	U:	i:	D
f	O:	h	Y	i:	G
f	T	h	Y:	i:	T
f	U	h	au	i:	_
f	U:	h	au:	i:	c
f	Y	h	ay	i:	c0
f	Y:	h	ay:	i:	f
f	_	h	c	i:	h
f	au	h	ey	i:	j
f	au:	h	ey:	i:	k
f	ay	h	i	i:	k0
f	ay:	h	i:	i:	l
f	c	h	k	i:	l0
f	c0	h	ou	i:	m
f	ey	h	ou:	i:	n
f	ey:	h	p	i:	n0
f	h	h	t	i:	p
f	i	h	yy	i:	p0
f	i:	i	C	i:	r
f	j	i	D	i:	r0
f	k	i	G	i:	s
f	k0	i	J	i:	t
f	l	i	J0	i:	t0
f	l0	i	N	i:	v
f	m	i	N0	i:	x
f	n	i	T	j	6y
f	n0	i	_	j	6y:
f	ou	i	c	j	9
f	ou:	i	c0	j	9:
f	oy	i	f	j	A
f	p	i	h	j	A:
f	p0	i	j	j	C
f	r	i	k	j	E
f	r0	i	k0	j	E:
f	s	i	l	j	I
f	t	i	l0	j	I:
f	t0	i	m	j	O
h	6y	i	m0	j	O:
h	6y:	i	n	j	T
h	9	i	n0	j	U
h	9:	i	p	j	U:
h	A	i	p0	j	Y
h	A:	i	r	j	Y:

j	_	k	f	l	T
j	au	k	h	l	U
j	au:	k	i	l	U:
j	ay	k	k0	l	Y
j	ay:	k	l	l	Y:
j	c	k	l0	l	au
j	c0	k	m	l	au:
j	ey	k	n	l	ay
j	ey:	k	n0	l	ay:
j	f	k	ou	l	c
j	h	k	ou:	l	c0
j	i	k	oy	l	ey
j	i:	k	p	l	ey:
j	j	k	p0	l	f
j	k	k	r0	l	i
j	k0	k	s	l	i:
j	l	k	t	l	j
j	l0	k	t0	l	k
j	m	k	v	l	k0
j	n	k0	6y	l	l0
j	n0	k0	6y:	l	m
j	ou	k0	9	l	n
j	ou:	k0	9:	l	n0
j	oy	k0	A	l	ou
j	p	k0	A:	l	ou:
j	p0	k0	O	l	oy
j	r	k0	O:	l	p
j	r0	k0	Y	l	p0
j	s	k0	Y:	l	r
j	t	k0	au	l	r0
j	t0	k0	au:	l	s
j	v	k0	l	l	t
k	6y	k0	n	l	t0
k	6y:	k0	ou	l	v
k	9	k0	ou:	l	yy
k	9:	k0	oy	l0	6y
k	A	k0	v	l0	6y:
k	A:	l	6y	l0	9
k	O	l	6y:	l0	9:
k	O:	l	9	l0	A
k	T	l	9:	l0	A:
k	U	l	A	l0	C
k	U	l	A:	l0	E
k	U:	l	C	l0	E:
k	U:	l	E	l0	I
k	Y	l	E:	l0	I:
k	Y:	l	I	l0	O
k	_	l	I:	l0	O:
k	au	l	O	l0	T
k	au:	l	O:	l0	U

l0	U:	m	Y:	n	U:
l0	Y	m	_	n	Y
l0	Y:	m	au	n	Y:
l0	_	m	au:	n	_
l0	au	m	ay	n	au
l0	au:	m	ay:	n	au:
l0	ay	m	c	n	ay
l0	ay:	m	c0	n	ay:
l0	c	m	ey	n	c
l0	c0	m	ey:	n	c0
l0	ey	m	f	n	ey
l0	ey:	m	h	n	ey:
l0	f	m	i	n	f
l0	h	m	i:	n	h
l0	i	m	j	n	i
l0	i:	m	k	n	i:
l0	j	m	k0	n	j
l0	k	m	l	n	k0
l0	k0	m	l0	n	l
l0	m	m	n	n	l0
l0	m0	m	n0	n	m
l0	n	m	ou	n	n0
l0	n0	m	ou:	n	ou
l0	ou	m	oy	n	ou:
l0	ou:	m	p	n	oy
l0	p	m	p0	n	p
l0	p0	m	r	n	p0
l0	r	m	r0	n	r
l0	r0	m	s	n	r0
l0	s	m	t	n	s
l0	t	m	t0	n	t
l0	t0	m	v	n	t0
l0	v	m0	k	n	v
m	6y	m0	p	n	yy
m	6y:	m0	t	n0	6y
m	9	n	6y	n0	6y:
m	9:	n	6y:	n0	9
m	A	n	9	n0	9:
m	A:	n	9:	n0	A
m	C	n	A	n0	A:
m	E	n	A:	n0	C
m	E:	n	C	n0	E
m	I	n	E	n0	E:
m	I:	n	E:	n0	I
m	O	n	I	n0	I:
m	O:	n	I:	n0	O
m	T	n	O	n0	O:
m	U	n	O:	n0	T
m	U:	n	T	n0	U
m	Y	n	U	n0	U:

n0	Y	ou	T	ou:	Y
n0	Y:	ou	U	ou:	_
n0	_	ou	U:	ou:	au
n0	au	ou	Y	ou:	ay
n0	au:	ou	Y:	ou:	c
n0	ay	ou	_	ou:	c0
n0	ay:	ou	au	ou:	ey
n0	c	ou	au:	ou:	f
n0	c0	ou	ay	ou:	h
n0	ey	ou	ay:	ou:	i
n0	ey:	ou	c	ou:	j
n0	f	ou	c0	ou:	k
n0	h	ou	ey	ou:	k0
n0	i	ou	ey:	ou:	l
n0	i:	ou	f	ou:	l0
n0	j	ou	h	ou:	m
n0	k	ou	i	ou:	n
n0	k0	ou	i:	ou:	n0
n0	l	ou	j	ou:	ou
n0	l0	ou	k	ou:	oy
n0	m	ou	k0	ou:	p
n0	ou	ou	l	ou:	p0
n0	ou:	ou	l0	ou:	r
n0	p	ou	m	ou:	r0
n0	p0	ou	m0	ou:	s
n0	r	ou	n	ou:	t
n0	r0	ou	n0	ou:	t0
n0	s	ou	ou	ou:	v
n0	t	ou	ou:	ou:	x
n0	t0	ou	oy	oy	j
n0	v	ou	p	p	6y
ou	6y	ou	p0	p	6y:
ou	6y:	ou	r	p	9
ou	9	ou	r0	p	9:
ou	9:	ou	s	p	A
ou	A	ou	t	p	A:
ou	A:	ou	t0	p	C
ou	C	ou	v	p	E
ou	D	ou	x	p	E:
ou	E	ou:	6y	p	I
ou	E:	ou:	9	p	I:
ou	G	ou:	A	p	O
ou	I	ou:	C	p	O:
ou	I:	ou:	D	p	T
ou	J	ou:	E	p	U
ou	J0	ou:	G	p	U
ou	N	ou:	I	p	U:
ou	N0	ou:	O	p	U:
ou	O	ou:	T	p	Y
ou	O:	ou:	U	p	Y:

p	_	p0	i:	r	yy
p	au	p0	l	r0	6y
p	au:	p0	ou	r0	6y:
p	ay	p0	ou:	r0	9
p	ay:	r	6y	r0	9:
p	c	r	6y:	r0	A
p	c0	r	9	r0	A:
p	ey	r	9:	r0	C
p	ey:	r	A	r0	E
p	f	r	A:	r0	E:
p	h	r	C	r0	I
p	i	r	D	r0	I:
p	i:	r	E	r0	O
p	j	r	E:	r0	O:
p	k	r	I	r0	T
p	k0	r	I:	r0	U
p	l	r	O	r0	U:
p	l0	r	O:	r0	Y
p	m	r	T	r0	Y:
p	n	r	U	r0	_
p	n0	r	U:	r0	au
p	ou	r	Y	r0	au:
p	ou:	r	Y:	r0	ay
p	oy	r	au	r0	ay:
p	r0	r	au:	r0	c
p	s	r	ay	r0	c0
p	t	r	ay:	r0	ey
p	t0	r	c	r0	ey:
p	v	r	c0	r0	f
p0	6y	r	ey	r0	i
p0	6y:	r	ey:	r0	i:
p0	9	r	f	r0	j
p0	9:	r	h	r0	k
p0	A	r	i	r0	k0
p0	A:	r	i:	r0	m0
p0	E	r	j	r0	n0
p0	E:	r	k	r0	ou
p0	I	r	k0	r0	ou:
p0	I:	r	l	r0	p
p0	O	r	l0	r0	s
p0	O:	r	m	r0	t
p0	Y	r	n	r0	t0
p0	Y:	r	n0	s	6y
p0	au	r	ou	s	6y:
p0	au:	r	ou:	s	9
p0	ay	r	p	s	9:
p0	ay:	r	p0	s	A
p0	ey	r	t	s	A:
p0	ey:	r	t0	s	C
p0	i	r	v	s	E

s	E:	t	I:	t0	O:
s	I	t	O	t0	Y
s	I:	t	O:	t0	Y:
s	O	t	T	t0	au
s	O:	t	U	t0	au:
s	T	t	U	t0	ay
s	U	t	U:	t0	ay:
s	U:	t	U:	t0	c
s	Y	t	Y	t0	ey
s	Y:	t	Y:	t0	ey:
s	_	t	_	t0	i
s	au	t	au	t0	i:
s	au:	t	au:	t0	j
s	ay	t	ay	t0	ou
s	ay:	t	ay:	t0	ou:
s	c	t	c0	t0	oy
s	c	t	ey	t0	v
s	c0	t	ey:	t0	yy
s	ey	t	f	v	6y
s	ey:	t	h	v	9
s	f	t	i	v	9:
s	h	t	i:	v	A
s	i	t	j	v	A:
s	i:	t	k	v	D
s	j	t	k0	v	E
s	k	t	l	v	E:
s	k0	t	l0	v	I
s	l0	t	m	v	I:
s	m	t	n	v	O
s	n0	t	n0	v	O:
s	ou	t	ou	v	U
s	ou:	t	ou:	v	U:
s	oy	t	oy	v	Y
s	p	t	p	v	Y:
s	p0	t	p0	v	au
s	r	t	r0	v	au:
s	r0	t	s	v	ay
s	t	t	v	v	ay:
s	t0	t	yy	v	c
s	v	t0	6y	v	c0
t	6y	t0	6y:	v	ey
t	6y:	t0	9	v	ey:
t	9	t0	9:	v	i
t	9:	t0	A	v	i:
t	A	t0	A:	v	j
t	A:	t0	E	v	k
t	C	t0	E:	v	k0
t	E	t0	I	v	l
t	E:	t0	I:	v	m
t	I	t0	O	v	n

v	ou	v	t0	x	l0
v	ou:	x	C	x	n0
v	oy	x	T	x	p
v	p	x	_	x	p0
v	p0	x	c0	x	r0
v	r	x	f	x	s
v	r0	x	h	x	t
v	s	x	k	x	t0
v	t	x	k0	yy	j